STATISTICAL MATHEMATICS

UNIVERSITY MATHEMATICAL TEXTS

GENERAL EDITORS

ALEXANDER C. AITKEN, D.Sc., F.R.S.
DANIEL E. RUTHERFORD, Dr. Math.

Determinants and Matrices	A. C. Aitken, D.Sc., F.R.S.
Statistical Mathematics	A. C. Aitken, D.Sc., F.R.S.
Waves	C. A. Coulson, Ph.D.
Integration	R. P. Gillespie, Ph.D.
Infinite Series	J. M. Hyslop, D.Sc.
Integration of Ordinary Differential Equations	E. L. Ince, D.Sc.
Analytical Geometry of Three Dimensions	Prof. W. H. McCrea, Ph.D.
Functions of a Complex Variable	E. G. Phillips, M.A., M.Sc.
Vector Methods	D. E. Rutherford, Dr. Math.
Theory of Equations	Prof. H. W. Turnbull, F.R.S.

Other volumes in preparation

STATISTICAL MATHEMATICS

BY

A. C. AITKEN, M.A., D.Sc., F.R.S.

READER IN STATISTICS AND ACTUARIAL MATHEMATICS
AT EDINBURGH UNIVERSITY

OLIVER AND BOYD
EDINBURGH AND LONDON
NEW YORK: INTERSCIENCE PUBLISHERS, INC.
1944

First Edition . . 1939
Second Edition . . 1942
Third Edition . . 1944

PRINTED IN GREAT BRITAIN BY
OLIVER AND BOYD LTD., EDINBURGH

CONTENTS

CHAPTER I

STATISTICS AS A SCIENCE: AXIOMS OF PROBABILITY

CHAPTER II

PROBABILITY AND FREQUENCY DISTRIBUTIONS: GRAPHICAL REPRESENTATION: CALCULATION OF MOMENTS

CHAPTER III

SPECIAL PROBABILITY DISTRIBUTIONS

CHAPTER IV

PRACTICAL CURVE FITTING WITH STANDARD CURVES

CHAPTER V

PROBABILITY AND FREQUENCY IN TWO VARIATES

CHAPTER VI

THE METHOD OF LEAST SQUARES: MULTIVARIATE CORRELATION: POLYNOMIAL AND HARMONIC REGRESSION

CHAPTER VII

PROBABILITY DISTRIBUTIONS OF STATISTICAL COEFFICIENTS

APPENDIX

CHAPTER I

STATISTICS AS A SCIENCE: AXIOMS OF PROBABILITY

1. Introductory. The word "statistics" is defined in the *Concise Oxford Dictionary* as follows: in the plural, "numerical facts systematically collected, as statistics of population, crime"; in the singular, "science of collecting, classifying and using statistics." This definition adequately conveys the present meaning of the word; but the term was once restricted, as its derivation shows, to systematic collections of data descriptive of political communities, a domain partly taken over now by the more special word "demography."

The word *statistics* (in the plural) is used nowadays to characterize "numerical facts systematically collected" in any field whatever of observation or experiment. The technique of collecting data and the principles to be heeded in order to avoid bias in the interpretation are described at length and exemplified in chapters of more extensive treatises which the reader may consult. He may also form a general idea of practical details by studying the prefatory description of method in some actual published investigation, for example into housing and economic conditions in a particular town or area. In any case the principles to be observed in arranging a statistical investigation can be thoroughly grasped only when the analysis used to interpret the data is well understood; and this involves a knowledge of the *science* of statistics (in the singular).

The intermediate stage of *tabulation*, by which collected

data are set out in the most perspicuous form for analysis or inspection with a particular aim, is also usually the subject of a chapter, with illustrative examples and criticisms, in larger treatises than the present one. Here again the reader may learn much from the attentive perusal of statistical year-books and similar publications, and from the results tabulated in other published investigations. The principles are those of logical classification of different categories; and the art of tabulation rests in making the relation of the categories and the numbers in various categories as clear as possible to the eye yet compact on the printed page. Thus one may have statistics of employed persons according to age, sex, district, trade and wage; how can the respective numbers best be set out in one or more tables with rows and columns, row-totals, column-totals, sub-totals and grand totals? This is a typical problem of tabulation, and the chief aids towards resolving it rest on experience and common sense.

Statistics involves classification by number in categories. Let us note for further reference the possible relations of individuals in two categories A and B. It may be that an individual of the collection cannot be both A and B at the same time; for example if a coin falls "heads," it certainly has not fallen "tails." The categories A and B are then mutually exclusive; their relation is that of "either ... or." On the other hand, the categories A and B may be of such a kind that an individual may belong to both at the same time; the relation of such categories is that of "both ... and."

2. Statistics as a Science. The concern of the present book will for the most part be with statistics (in the singular) as a science. The typical order of development of the "exact" sciences (as they are somewhat loosely called) has been along the following lines. First of all, the examination of data collected in a particular field of inquiry is found to disclose elements of regularity,

suggesting a law or laws. This is the stage of *inductive synthesis*. These laws are expressed, if possible, in the form of logical or numerical axioms, resembling those of Euclidean geometry. The methods of logic and mathematics are then brought into play to develop the consequences of the axioms, producing an assemblage of theorems or propositions. This department of the science, namely the posing of axioms and the deduction of theorems, is usually called the *pure* branch of the science. Even if future observations should invalidate the axioms extrinsically, the discrepancies between theory and fact being too great to be explained away, these axioms and the deductions based on them would still have an abstract validity, as a logical structure of propositions exempt from self-contradiction; but for the description and explanation of the phenomena a new set of axioms would have to be found. On the other side, the corroborative part of the science consists in interpreting the abstract functions, formulæ, equations, constants, invariants and the like, which occur in the pure formulation, as measures and measurable relations of actual phenomena, or numbers constructed from those measures in a definite way. This interpretative aspect constitutes the *applied* branch of the science.

Such a division or dichotomy into pure and applied can be recognized in almost any science. A good example is Newtonian dynamics, according to which the motions of all bodies in the universe were presumed to obey certain axioms and postulates, namely Newton's laws of force and motion and the law of gravitation. Later experiments, more numerous, more delicate, more comprehensive, suggested that this formulation, though describing almost all observed dynamical phenomena with a precision unprecedented in history, did not sufficiently account for certain exceptional facts, such as the precession of the perihelion of Mercury. The discrepancies between prediction and actuality were extraordinarily small, but they were persistent. There thus

arose a theory, or rather a succession of supplementary theories, of relativity, formulated on a new axiomatic basis by which the discrepancies of the earlier one might be reconciled, or removed. This reformulation of hypotheses still proceeds, is still incomplete, and undergoes modification from time to time.

What is the axiomatic basis of the science of statistics, and what are the facts upon which the inductive synthesis is based ? The facts are certain regularities which have been observed in the *proportionate frequency* with which certain simple events happen or do not happen, when the circumstances under which they may occur are reconstructed again and again *in repeated trials* ; and the axioms, and the structure of theorems founded upon them, constitute the subject called *mathematical probability*. As for the facts, anyone who is interested can collect a few for himself. Spin an ordinary coin a large number of times, and one can hardly fail to notice that the proportions of heads and of tails are very nearly equal ; or shake a well-made die repeatedly from a dice-box and one will find that after many trials each face of the die has turned up in about one-sixth of the total number of trials.

Example. The reader is recommended to experiment with simple repeated trials of this kind, and for future reference to record the results in sequence, in the order in which they occur. For example, the record of spins of a coin might be

00101 01110 01101 00001 10111 ...

or the like, where " 1 " denotes " heads," and " 0 " " tails."

It is instinctive to look for some cause for this approximate equality of frequency in heads and tails, and natural to locate this cause as somehow resident in the two-sided nature and appreciable symmetry of the coin ; or to ascribe the approximate equality of frequency of the faces of the die to its six-sided and nearly uniform configuration. Simple ideas such as these suggest by generalisation and abstraction the axioms of probability ;

but the choice of axioms may be made in various ways, which lead to different formulations of the theory of probability.

3. Survey of Various Definitions of Probability. No single particular definition of probability has so far met with predominating acceptance. The requisites of a satisfactory basis would be these: breadth of application, sufficient closeness to the intuitions in which the concept originates, and freedom from excessive complexity or abstruseness. No theory as yet proposed has been able to make these requisites compatible. We may survey some contrasting standpoints.

Probability as the Logic of Uncertain Inference. One view is that probability may be regarded as a kind of extension of classical logic, an extension conveniently described as the "logic of uncertain inference." This view has been expounded by J. M. Keynes in *A Treatise on Probability* (London, 1921), especially in Part II, Chapters X-XVII, where references to earlier expositions are given. Probability is here regarded as "the degree of our rational belief" in the truth of a given proposition, such belief being contingent on a body of relevant knowledge. A logical algebra is developed, but the theorems are stated in symbolic, not in numerical or metrical terms, and can be applied to the objective problems of statistics only by an abrupt and dubious transition from the symbolic to the metrical.

Probability *à Priori*, and Probability as Relative Frequency. As our simple illustrations of the coin and the die have suggested, the crude intuition of probability rests on the observation that when a given set of circumstances S, such as a symmetrical coin spun rapidly, has been present on numerous occasions in the past, it has been associated in a nearly constant proportion of those occasions with some event E, such as the fall of "heads."

The *à priorist* theory directs attention to the set of

circumstances S, or rather to the invariant part of S. In many spins of a coin or die something remains unchanged, namely those properties which describe the coin or die as a rigid constant configuration. The *à priorist* will regard the probabilities of falls 1, 2, 3, 4, 5, 6 of a die as some part of the description of the die, as measuring indeed some quality resident in the structure of the die, before any spinning is performed. Now the classical *à priori* definition took account only of a very limited class of " systems " S, namely those possessing *symmetry*, in the sense that the different aspects (such as faces 1, 2, 3, 4, 5, 6 of the die) were presumed physically indistinguishable. Such an assumption is an idealization of the facts, for we can never hope to test completely the symmetry of any actual coin or die; not only would the tests be infinitely many and impossibly delicate, but the concept of the rigidity and permanence in time of a material body is not sustained by modern physics. However, symmetry being presumed, the six faces 1, 2, 3, 4, 5, 6 were characterized as " equally likely " to be found uppermost after any throw, and the probability of 1/6 was attributed to each of these " events." More generally, if n equally likely aspects of a proposed system S were discriminated, m of these being favourable to the event E, the probability of E with respect to S was defined as $p(E\,;S) = m/n$.

Criticism is easy. The logician will not fail to pounce upon the words " equally likely," pointing out that they are synonymous with " equally probable," and that therefore probability is being defined by what is probable, a *circulus in definiendo* being thus committed. Postponing the defence, we may pass on to inquire what could be the definition of probability, should the tests have disclosed asymmetry in S. The inquiry is most pertinent, for the heterogeneous and the asymmetrical are the prevalent order of nature, the homogeneous and the symmetrical being the exception. One has no difficulty for example in

conceiving a die which might be an irregular hexahedron, heterogeneous in density and with non-parallel and unequal opposite edges and faces. Such dice, and more complicated asymmetrical systems, have been subjected to repeated trials, which have shown a tendency of relative frequency of falls towards a constancy resembling that observed in symmetrical systems.

Stability of Relative Frequency. Another view from the angle of " common sense," in some respects antithetical to the view just mentioned, is the frequency view. Here the invariability of the configurative part of S, whether symmetrical or unsymmetrical, is tacitly assumed, and attention is concentrated upon the sequence of trials, and the incidence of E in these. For example, the die is thrown again and again. When E occurs, let us write 1 ; when E does not occur, let us write 0. A succession of n trials then gives a sequence

$$\mathrm{A} = a_1\, a_2\, a_3\, a_4 \ldots a_n, \qquad (1)$$

each a_j being 1 or 0.

Let m be the number of 1's in this sequence. A very limited experience, such as spinning a coin or die 10 times on several occasions, will show that in a finite number n of trials made upon the same system S on two or more occasions, different values of m are not only possible but usual. Thus, if E is the throw of an ace with a single die, 100 throws may on one occasion give $m = 15$ and on another occasion give $m = 20$. It follows that in order to define a probability $p(E\,;S)$ which shall be unique and not discordant with experience, we must idealize once again, postulating a *limiting process* as n tends to infinity and writing

$$\lim_{n\to\infty} m/n = p(E\,;S). \quad . \quad . \quad . \quad (2)$$

This is in fact a definition, supported by a certain school of statisticians, based upon the limit of frequency ratio or relative frequency m/n. Though at first sight attractive,

it fades a little on scrutiny. Granted the postulate of this limit p for one sequence of trials upon S, can we accept the more stringent postulate that the same limiting value p is obtained for any other infinite sequence of trials on S? Not without further assumptions, for one might imagine a mechanism sufficiently delicate to throw heads with a coin, or an ace with a die, on almost all occasions. There is therefore some restriction on the manner of throwing, or on the initial state of S. This restriction is usually stated in the form of a condition that successive throws must be "random," but this merely transfers the burden of explanation to a new and undefined concept, "randomness." To discuss various attempts to define randomness would take us too far afield. It is easy to say that randomness is absence of any law; but what is "law" in this connexion?

Another difficulty is that the tendency of relative frequency m/n towards a limit p is different in nature from the corresponding tendency to a limit which mathematicians have discerned and used in the infinite sequences of mathematical analysis. To take a classical example, in the sequence defining a certain simple geometric series,

$$1,\ 1-\tfrac{1}{2},\ 1-\tfrac{1}{2}+\tfrac{1}{4},\ 1-\tfrac{1}{2}+\tfrac{1}{4}-\tfrac{1}{8}, \qquad (3)$$

the deviations of the successive terms from $\frac{2}{3}$ are respectively $\frac{1}{3}$, $-\frac{1}{6}$, $\frac{1}{12}$, $-\frac{1}{24}$, ..., each being numerically half its predecessor, so that, given a small number ϵ, such as 1/1000000, we can always find some term sufficiently far along the sequence, after and including which *all* terms deviate from $\frac{2}{3}$ by less than ϵ. Thus $\frac{2}{3}$ is the limit of this sequence. But what can be asserted concerning the sign and magnitude of the deviation ϵ_n, considered as a function of n, in

$$\epsilon_n = m/n - p(E : S)\ ?$$

It would seem that the only kind of assertion about ϵ_n which would carry conviction would itself involve some-

where the notion of probability; and here the risk of committing a circle in definition again raises its head.

It should be added that the chief defects of the approach to probability by limit of frequency ratio have lately been removed by the work of de Mises, Copeland, Dörge, Wald and others. These writers admit only certain sequences A of suitable postulated properties, including that of limiting ratio; but some logical difficulties remain, and the modified formulations lose the primitive simplicity in which they originated.

It would seem, however, that a more natural course, and one more in line with the general method of science, would be to try to explain the *effect*, namely the relative frequency of E, by an analysis of the *cause*, namely the system S. This suggests a return to the *à priori* standpoint; and it may be noted that several authors at the present time, Fréchet, Kolmogoroff, Cramér and others, have been independently engaged in rehabilitating the *à priori* definition by furnishing it with a better axiomatic basis.

4. Probability as Measure of a Sub-Aggregate. Let us examine more closely the system S, keeping some simple system such as a coin or die in mind. The approximately constant element in our sequences A, namely the almost stable frequency ratio of E, must reflect—at least so our intuition suggests—the constant element of S, such as the rigid configuration of a coin or die; the irregularity which we name *randomness* doubtless reflects the variable part of S, such as the initial position, velocity and angular velocity of projection.

What is S when an unsymmetrical and heterogeneous die is spun and falls? It consists of (i) the die, specified as a particular constant rigid body, (ii) the floor or table on which it may impinge or finally rest, (iii) the surrounding air, and so on; together with (iv) the circumstances of projection, described by coordinates of initial position,

momentum and angular momentum. The coordinates specifying the rigidity of the die and the configuration of the table or floor are constant components of S, the other initial coordinates of S are variable. The set of coordinates of S at the instant of projection may be called the initial phase. Each variable coordinate, such as the initial position, or the initial momentum, has a certain field of variation. Hence we must assume a *set of possible phases* which, if they can be enumerated in some order, may be designated by $S_1, S_2, \ldots, S_i, \ldots$; and this ensemble of possible initial phases S_i constitutes an aggregate S of the kind specially studied in pure mathematics.* If dynamical determinism be assumed, but not otherwise, the initial phase will decide whether or not the event E will occur. Consequently the possible initial phases may be classified as E-phases or not-E-phases (let us say $\bar{E}$-phases), so that the whole phase aggregate is divided into two sub-aggregates. Now the question of assigning a *measure* to such aggregates has been deeply studied in modern pure mathematics, the guiding idea being that of extending as widely as possible the scope of a concept familiar in simple cases, namely the cardinal number of a finite set of objects, the length of a line, the area of a surface, the volume of a solid. If M is the measure of the whole aggregate S of possible phases, and pM the measure of the aggregate of E-phases contained in it, then p is the probability $p(E\,;\,S)$.

Something has been glossed over here; there is the tacit assumption that the initial phases are "equally likely." But let us insist that the question of equal likeliness is not one for the abstract formulation at all; for to specify the aggregate is in effect to say that its elements, the initial phases, are equally likely. For example, if the aggregate were of points on a continuous line segment, and the measure were ordinary length, then

* We use the same letter S as before, regarding the system now as the totality of its possible phases.

we have implied in this description that all points in the segment are equally likely. On the other hand, the question of equal likeliness is crucial in the *application* to experiment or observation, that is, in applied statistics, where a wrong choice of the aggregate may alter all the probabilities. This has long been known in problems of so-called geometrical probability. For example, given a circle, let a chord be drawn across it at random: what is the probability that the length of the chord exceeds half the diameter? It depends entirely on the manner in which the chord is drawn. If it is done by taking a point on the circumference and then drawing the chord at any angle, all angles being thus supposed equally likely, then the probability is 2/3; but if it is done by taking any diameter and drawing the chord at right angles to any point taken in the diameter, the diameters and points being equally likely, then the probability is $\sqrt{3}/2$.

The inclusion of the words "equally likely" in a definition is in fact a concession; it puts the reader more gently at terms with the abstract formulation by anticipating its chief future application. The usage is not uncommon. When a point is defined as "that which has position but no magnitude" the same appeal is made to an application, but the same suspicion of a circle in definition is incurred, for how can position be defined without the notion of a point? And if a straight line is defined as "lying evenly" between its extreme points, what else does "evenly" mean but "in a straight line"? Every definition which is not pure abstraction must appeal somewhere to intuition or experience by using some such verbal counter as "point," "straight line" or "equally likely," under the stigma of seeming to commit a circle in definition.

This prologue, though it has omitted many subtler points which could be amplified at very great length, must now be cut short. To summarize: (i) events E are conceived as associated with, or caused by, phases S_i of circumstances; (ii) each S_i gives rise unambiguously

either to E or to $\bar{E}$; (iii) the phases S_i form in their totality a set or aggregate S, of which the phases favourable to E, and those favourable to $\bar{E}$, form complementary subsets; (iv) a measure M can be given to the whole set S, and if pM is the measure of the subset favourable to E, then p is the probability $p(E\,;S)$ of E with respect to S; (v) the question of equal likeliness of phases is the same as the question of specifying the aggregate and its measure, and in practical applications this must be determined by the circumstances of the particular problem. Let us finally add that the word phase can be extended to include coordinates other than dynamical ones; also that the name "fundamental probability set" is used by some writers for the set S of phases S_i.

5. Definition of Probability. In an elementary treatment a rigorous formulation in terms of general aggregates is not possible. It will be necessary to restrict consideration to aggregates with a finite number of elements only; in this case the measure of an aggregate or sub-aggregate is simply the number of elements it contains. The reader may take it that the theorems can be extended to more general aggregates.

Definition. If an event E can result from the phases of a system S, there being n different phases and no more, all equally likely *à priori*; and if m of these phases entail the occurrence of E (so that $n-m$ do not), then m/n is the probability $p(E\,;S)$ of E with respect to S.

Continuous Case. If the event E is described by the value of a continuous variable x, we may denote the probability that x is found between $x+\frac{1}{2}\Delta x$ and $x-\frac{1}{2}\Delta x$ by

$$p(x+\tfrac{1}{2}\Delta x,\, x-\tfrac{1}{2}\Delta x\,;S) \equiv \Delta p(x\,;S), \quad . \quad . \quad (1)$$

let us say. By supposing n to tend to infinity and Δx to tend to zero we reach the conception of a differential element of probability, or probability differential,

$$p(x+\tfrac{1}{2}dx,\, x-\tfrac{1}{2}dx\,;S) \equiv dp(x\,;S), \quad . \quad . \quad (2)$$

which, when no misunderstanding about S is likely to arise, we shall often denote briefly by dp.

Complementary Event. The failure of E is denoted by $\bar{E}$, and is called the *complementary event*. The probability of $\bar{E}$ is $(n-m)/n$, namely $1-p$ in the finite case, and likewise in the continuous case. This is often termed the *complementary probability* and denoted by q, so that $p+q=1$.

If n is finite and if E must inevitably happen in all of the n ways, then $p=1$ and E is "certain," while $q=0$ and $\bar{E}$ is "impossible." If, however, the system S depends on a non-finite set or results in events expressible by a continuous variable, we must not suppose that $p=1$ implies certainty, or $p=0$ impossibility. For example, if a point is taken on a line segment, the chance of a particular point P being taken is 0; but *some* point is taken, and so the point P cannot be regarded as impossible.

6. Addition and Multiplication of Probabilities. Dependent and Independent Events. An event F will be said to be *dependent* on an event E when the happening of either E or $\bar{E}$ alters the probability of F; and in the contrary case F will be said to be *independent* of E. An extreme case of dependence is that in which the happening of either E or F makes the probability of the other equal to zero. The events are then said to be *mutually exclusive*. (In the continuous case we must take cognizance of "almost mutually exclusive" and "almost independent" events, just as we have of "almost impossible" events for which $p=0$.)

The addition theorem of probability is applicable to events which are mutually or almost mutually exclusive.

Theorem. When an event E may happen in the form of any one of r mutually exclusive events E_j, $j=1, 2, 3, \ldots r$, in a system S which has n equally likely phases, the probability of E_j being p_j, then the probability of E is

$$p(E\,;S)=p_1+p_2+\ldots+p_r=\Sigma p_j \quad . \quad . \quad (1)$$

Proof. If n_j of the n phases entail E_j then $p_j = n_j/n$. Since the phases do not overlap (otherwise the events E_j would not be mutually exclusive) the total number of phases entailing one or other of the E_j is $\sum_j n_j$; and so

$$p(E\,;S) = \sum_j n_j/n = \sum_j p_j.$$

The theorem, which is sometimes called the theorem of *Total Probability*, continues to hold for systems expressed by a non-finite n or by a continuous variable.

The multiplication theorem, or theorem of *Compound Probability*, refers in the first instance to independent events, but can easily be made applicable, with a suitable definition of conditioned probability for dependent events, to the latter case.

Theorem. If E_j, $j = 1, 2, 3, \ldots, r$, are r independent events, each with respect to its own system S_j, the probability that they all happen when all the S_j are in operation is

$$p(E\,;S) = p_1 p_2 \ldots p_r, \quad . \quad . \quad . \quad (2)$$

where E denotes the compound event consisting in the happening of all the E_j, S denotes the compound system consisting in the operation of all the S_j, and $p_j = p(E_j\,;S_j)$.

Proof. Let n_j denote the number of phases of S_j, and of these let m_j entail E_j. Now each of the n_j phases of S_j may be paired in turn with each of the n_k phases of S_k, giving rise to $n_j n_k$ compound phases of the double system (S_j, S_k). By similar reasoning the m_j phases entailing E_j may be paired in turn with the m_k phases entailing E_k, giving rise to $m_j m_k$ compound phases of (S_j, S_k) entailing the double event (E_j, E_k).

By similar reasoning, or step by step, there are altogether $n_1 n_2 \ldots n_r$ phases of the compound system $(S_1, S_2, \ldots, S_r) = S$, and of these $m_1 m_2 \ldots m_r$ entail the compound event $(E_1, E_2, \ldots, E_r) = E$. Hence the probability of E with respect to S is

$$p(E\,;\,S) = m_1 m_2 \dots m_r / n_1 n_2 \dots n_r$$
$$= p_1 p_2 \dots p_r.$$

Once again we must content ourselves with the statement that the theorem remains true for independent or "almost independent" systems involving infinite aggregates or continuous variables.

By modifying the definition of $p_2, p_3, \dots, p_r$ we may prove an analogous theorem for a chain of events $E_1, E_2, \dots, E_r$, each of which influences the probability of its successors.

Let $p_2 \equiv p(E_2\,;\,E_1, S_2)$ denote the probability of E_2 after E_1 has happened, $p_3 \equiv p(E_3\,;\,E_1, E_2, S_3)$ denote the probability of E_3 after E_1 and E_2 have happened, and so on. Slight consideration will show that this simply involves putting the events *in an order of time* and that then, with the new interpretation of $p_2, p_3, \dots, p_r$, the above proof proceeds exactly as before. Hence we have the theorem of compound probability for a chain of conditioned events:

$$p(E\,;\,S)$$
$$= p(E_1)p(E_2\,;\,E_1)p(E_3\,;\,E_1, E_2) \dots p(E_r\,;\,E_1, E_2, \dots, E_{r-1}). \quad (3)$$

These theorems of addition and multiplication of probabilities are the fundamentals upon which the mathematical theory of statistics is raised. Since addition and multiplication are operations of ordinary algebra, we may anticipate that there is an algebra of probability depending on these operations, according to which expressions representing independent systems S_j can be compounded in product and the resulting probabilities found by inspection of terms. This algebra is the algebra of *generating functions* of probability, which we shall consider from an elementary standpoint in the next section.

Ex. 1. The probability of throwing two consecutive aces with a true die is $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$.

Ex. 2. The probability of throwing a head and a tail with two coins is $\frac{1}{2}$.

Ex. 3. The probability of throwing a total of 8 points with two dice is 5/36. (The mutually exclusive events are 6+2, 5+3, 4+4, 3+5, 2+6.)

Ex. 4. A bag contains 4 black and 3 white balls. Show that the probability of drawing 3 black in succession is 64/343 if the ball drawn is replaced each time, 8/49 if the first ball drawn is replaced but not the others, 4/35 if no ball is replaced.

Ex. 5. The events E_1 and E_2 are neither independent nor mutually exclusive. Denote by p_{12} the probability that E_1 and E_2 both happen. Prove that the probability that at least one of E_1 and E_2 happen is $p_1+p_2-p_{12}$.

Ex. 6. Generalize the preceding theorem to r events $E_1, E_2, \ldots, E_r$. Prove, with an analogous notation, that the probability that at least one of the events happens is

$$\Sigma p_j - \Sigma\Sigma p_{jk} + \Sigma\Sigma\Sigma p_{jkl} - \ldots + (-)^r p_{12\ldots r}.$$

7. Generating Functions of Probability. We shall often denote the probability that a variable x takes a particular value x_j by $\phi(x_j)$, and we shall use the following nomenclature:

Probability Function. The function $\phi(x)$ is the *probability function.* When the set of values of x is continuous we shall write the *probability differential* $dp = \phi(x)dx$ for the probability that x is found in the range $(x-\frac{1}{2}dx,\ x+\frac{1}{2}dx)$. In this case $\phi(x)$ is often called the *probability density.*

Variate. A variable which has a probability function will be called a *variate.*

Generating Function. Associated with $\phi(x)$ we introduce the *generating function* (g.f.) of probability, defined by

$$G(t) \equiv G(t\,;\,\phi) = \Sigma\phi(x_j)\,t^{x_j} \quad . \quad . \quad . \quad (1)$$

for variables which take discrete values, and by

$$G(t) = \int \phi(x)t^x dx \quad . \quad . \quad . \quad (2)$$

for continuous variables, the integral being over the whole range of possible values of x.

Ex. 1. The generating function of probability for heads in a symmetrical coin is

$$\tfrac{1}{2}t^1+\tfrac{1}{2}t^0 = \tfrac{1}{2}(t+1).$$

Ex. 2. The g.f. for a symmetrical six-sided die is

$$\tfrac{1}{6}(t+t^2+t^3+t^4+t^5+t^6) = \tfrac{1}{6}t(1-t^6)/(1-t).$$

Ex. 3. The g.f. for an unsymmetrical coin in which the probability of heads is p, of tails q, is $pt+q$.

Ex. 4. Write down the g.f. for a symmetrical four-sided die ; also for an unsymmetrical one in which the probabilities of faces marked 1, 2, 3, 4 are p_1, p_2, p_3, p_4.

Ex. 5. If all points on the straight line from $x = 0$ to $x = 1$ are equally probable, the g.f. is

$$\int_0^1 t^x dx = (t-1)/\log_e t.$$

8. Properties of Generating Functions. Suppose first that we have an event E_1 and its complement $\bar{E}_1$, of respective probabilities p_1 and q_1, and a second independent event E_2 with its complement $\bar{E}_2$, of probabilities p_2 and q_2. Then the compound probabilities of the four mutually exclusive events

$$(E_1, E_2), (E_1, \bar{E}_2), (\bar{E}_1, E_2), (\bar{E}_1, \bar{E}_2) \quad . \quad . \quad (1)$$

are respectively p_1p_2, p_1q_2, q_1p_2, q_1q_2. Let us relate these to the terms on the right of the algebraic identity

$$(p_1t_1+q_1)(p_2t_2+q_2) = p_1p_2t_1t_2+p_1q_2t_1+q_1p_2t_2+q_1q_2. \quad (2)$$

Study of this identity will reveal the most important property of generating functions. The disjunction between *added* terms (those linked by plus signs), both in the factors on the left and in the expanded product on the right, reflects in each case the disjunction into a number of

mutually exclusive events. The operations of *multiplication*, on the other hand, are carried out on expressions symbolizing *independent* events. For example, the multiplication of the two factors on the left interprets the compounding of the two independent systems S_1 and S_2 of which they are the generating functions; and the results of multiplication visible in single terms on the right, such as $p_1p_2t_1t_2$, represent at the same time the compounded probabilities, p_1p_2, and the compounded events, t_1t_2 characterizing (E_1, E_2). In fact the algebraic operations are faithfully carrying out the consequences of the two basic theorems of probability. Mere inspection will convince us that this is true not only for binomial expressions compounded in product as above, but for multinomial expressions, as in the following example.

Ex. 1. Let the reader consider events E_1, E_2, E_3 of probabilities p_1, p_2, p_3 with respect to S, events E'_1, E'_2, E'_3, E'_4 with probabilities p'_1, p'_2, p'_3, p'_4 with respect to an independent system S', and examine the product

$$(p_1t_1+p_2t_2+p_3t_3)(p'_1t'_1+p'_2t'_2+p'_3t'_3+p'_4t'_4)$$

in relation to the 12 events of the compound system $T = (S, S')$.

Regarding a compound system (S, S') as a single system and introducing further independent systems one at a time, we may prove step by step that to find the respective probabilities of all the mutually exclusive events arising from the compounding of r independent systems, we must construct the product of r expressions of the kind exemplified above, and examine the individual terms of the expansion.

Ex. 2. In an expansion of three factors such a term as $p_1p'_1p''_2t_1t'_1t''_2$ would be interpreted as meaning that the compound event (E_1, E'_1, E''_2) has probability $p_1p''_1p'_2$.

The variables t_j and so on are introduced for the sole purpose of preventing the terms from being merged together; for when the p_j are explicit fractions such as $\frac{1}{2}$, $\frac{3}{8}$ and the like some such device is needed.

Now suppose the event E_j involves the *addition* of x_j points to a score, or the assumption by an *additive* variate x of an increment x_j. In such a case we represent E_j by t^{x_j} rather than by t_j, taking advantage of the fact that when expressions like $p_j t^{x_j}$ and $p_k' t^{x_k}$ are multiplied together we have by the law of indices $p_j p_k' t^{x_j + x_k}$, the probabilities being *multiplied* as they ought to be, and the increments x_j and x_k being *added* as they ought to be. With this understanding, the system under which x may assume values x_j with probabilities p_j, $j = 1, 2, \ldots, r$, is characterized by the expression

$$\sum_j p_j t^{x_j}. \qquad (3)$$

But this is merely the generating function $G(t)$ of the system, and so we infer the important theorem, for discrete variates in finite sets:

The g.f. of a compound of independent systems is the product of the g.f.'s of the separate systems.

By a limiting process, with due precautions on the functions concerned, this multiplicative law can be extended to g.f.'s involving continuous variables. Thus, if $G_1(t)$ is the g.f. of the variable x, and $G_2(t)$ of a statistically independent variable y, then $G_1(t)G_2(t)$ is the g.f. of $x+y$; and so for more than two variables.

Ex. 3. The probabilities of 3 heads, 2 heads, 1 head and no heads in a throw of three symmetrical coins (or three separate throws of one coin) are the coefficients of t^3, t^2, t and 1 in the expansion of $(\frac{1}{2}t+\frac{1}{2})^3$, namely $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, $\frac{1}{8}$ respectively. Verify this also by enumeration of cases. (Write H for head, T for tail; then the cases are HHH; HHT, HTH, THH; HTT, THT, TTH; TTT.)

Ex. 4. The corresponding probabilities when the coin is

unsymmetrical, with probability p for heads and q for tails, are the coefficients in the expansion of $(pt+q)^3$.

Ex. 5. The probabilities when the three coins are unsymmetrical are the coefficients in the expansion of $(p_1t+q_1)(p_2t+q_2)(p_3t+q_3)$.

Ex. 6. The probabilities of n, $n-1$, ..., 2, 1, 0 heads in n throws of an unsymmetrical coin are the coefficients of powers of t in the expansion of $(pt+q)^n$.

Ex. 7. Write down the corresponding g.f. for the simultaneous throw of n *different* unsymmetrical coins.

Ex. 8. A tetrahedral, a cubical and an octahedral die, all symmetrical, are thrown together, their faces being numbered in each case from 1 upwards. Show that the probabilities of totals 3, 4, ..., 18 are arrayed by coefficients in the expansion of

$$\tfrac{1}{4} \cdot \tfrac{1}{6} \cdot \tfrac{1}{8}(1-t^4)(1-t^6)(1-t^8)/(1-t)^3.$$

Ex. 9. A coin is thrown n times. Each time a head occurs, 2 is added to the score; each time a tail occurs, 1 is subtracted. The g.f. is

$$(\tfrac{1}{2}t^2+\tfrac{1}{2}t^{-1})^n = 2^{-n}t^{-n}(t^3+1)^n.$$

Ex. 10. Four tickets marked 00, 01, 10, 11 respectively are placed in a bag, and drawn one at a time, being replaced each time. Prove that the chance of drawing five times and obtaining ticket numbers summing to 23 is the coefficient of t^2u^3 in the expansion of $4^{-5}(1+t+u+tu)^5 = 4^{-5}(1+t)^5(1+u)^5$.

Find this coefficient, and verify the result by enumeration.

9. Moments and Moment Generating Functions. It is convenient to describe a probability function $\phi(x)$ by certain coefficients or parameters connected with it, such as moments, seminvariants and others later to be defined. The *moments* commonly employed are based on powers of x, and are defined by

$$\mu'_r = \Sigma x^r \phi(x) \quad \text{or} \quad \int x^r \phi(x) dx, \tag{1}$$

according as the variate is discrete or continuous. The summation or integration is over the whole range of

possible values of x. If the values which x can take are discrete and spaced at unit intervals (for example if x records the number of heads in n throws of a coin) it is mathematically preferable to use *factorial moments*, defined by

$$\mu'_{(r)} = \Sigma x^{(r)}\phi(x),$$

where $$x^{(r)} = x(x-1)(x-2)\ldots(x-r+1). \quad . \quad (2)$$

Note. The privilege often accorded to ordinary "power" moments is one of custom only; no special sanctity attaches to them.

Mathematical Expectation. If $f(x)$ is a function of x, and $\phi(x)$ is the probability function, or $\phi(x)dx$ the probability differential, then the sum or integral

$$\Sigma f(x)\phi(x) \text{ or } \int f(x)\phi(x)dx \quad . \quad . \quad . \quad (3)$$

is called the *mathematical expectation* of $f(x)$. It is often denoted by $Ef(x)$. The r^{th} moment is therefore the mathematical expectation of x^r.

Moment Generating Functions. If we put $t = e^{\alpha}$ in the g.f. of probability $G(t)$, we obtain

$$G(e^{\alpha}) = \sum_x \phi(x)e^{\alpha x} \text{ or } \int \phi(x)e^{\alpha x}dx \quad . \quad . \quad . \quad (4)$$
$$= 1 + \mu'_1\alpha + \mu'_2\alpha^2/2! + \mu'_3\alpha^3/3! + \ldots,$$

provided that the sum or integral converges over a range of α and that expansion of e^{α} and integration term by term is permissible. This function, which we shall denote by $M(\alpha)$, may be regarded as *generating* the moments μ'_r, in the sense that μ'_r is the coefficient of $\alpha^r/r!$ in $M(\alpha)$. Of course α, like t, is a variable introduced to facilitate manipulation, in fact to carry the moments. We shall call $M(\alpha)$ the *moment generating function* (m.g.f.) of x or of $\phi(x)$, or of the system in question.

Factorial Moment Generating Functions. When factorial moments are in question, we can construct a *factorial moment generating function* (f.m.g.f.) very simply from the probability g.f. by the substitution $t = 1+\alpha$. For then we have

$$G(1+\alpha) = \sum_x \phi(x)(1+\alpha)^x \quad . \quad . \quad . \quad . \quad (5)$$
$$= 1+\mu'_{(1)}\alpha+\mu'_{(2)}\alpha^2/2!+\mu'_{(3)}\alpha^3/3!+\ldots,$$

by expanding $(1+\alpha)^x$ by the binomial theorem and summing the resulting terms.

Example. The f.m.g.f. of the distribution characterized by $(pt+q)^n$ is $(1+p\alpha)^n$.

Note. The reader who is acquainted with more advanced mathematics may observe that for moment generating functions the substitution $t = e^{iu}$ instead of $t = e^{\alpha}$ has a certain advantage. It gives the modified m.g.f.

$$\int e^{iux}\phi(x)dx, \quad . \quad . \quad . \quad . \quad (6)$$

a *Fourier transform* of $\phi(x)$. The integrand and integral are bounded, and the reciprocal theorems of Fourier transforms are available.

10. Seminvariants and Seminvariant Generating Functions. If the *logarithm* of the moment generating function $M(\alpha)$ can be expanded as a convergent series in powers of α in the form

$$L(\alpha) = \log_e M(\alpha) = \lambda_1\alpha+\lambda_2\alpha^2/2!+\lambda_3\alpha^3/3!+\ldots, \quad . \quad (7)$$

then $L(\alpha)$ is defined to be the *seminvariant g.f.*, and the coefficients λ_r are called the *seminvariants* * of the function $\phi(x)$. Since m.g.f.'s are compounded in product, s.g.f.'s must be compounded *in sum*, whence the theorem:

When independent systems are compounded the r^{th} *seminvariants* λ_r *of the separate systems are added to form the* r^{th} *seminvariant of the compound system.*

This *additive property of seminvariants* is indeed the

* The word "cumulant," suggested by R. A. Fisher, is perhaps to be preferred, since "seminvariant" is already appropriated in the theory of algebraic invariants.

reason for introducing them. In the same way, by taking the logarithm of the f.m.g.f. we can define a factorial s.g.f. and *factorial seminvariants.*

Example. The factorial seminvariants corresponding to $(pt+q)^n$ are np, $-np^2$, $2!np^3$, $-3!np^4$ and so on.

11. Change of Origin and Scale in Generating Functions. Change of Origin. If the origin from which the variate x is measured is transferred from $x = 0$ to $x = a$, any value x will be changed to $x-a$. Hence every factor t^x in a term of the probability g.f. will become t^{x-a}; but the accompanying probability $\phi(x)$, though changed in notation, will not be changed in value. Hence the effect is to multiply the whole g.f. by t^{-a}.

This very simple rule leads to corresponding ones for the m.g.f., f.m.g.f. and s.g.f., namely :

A change of origin from $\mathrm{x} = 0$ *to* $\mathrm{x} = \mathrm{a}$ *has the effect of multiplying the m.g.f. by* $\mathrm{e}^{-\mathrm{a}\alpha}$*; of multiplying the f.m.g.f. by* $(1+\alpha)^{-\mathrm{a}}$*; and of adding to the s.g.f. the term* $-\mathrm{a}\alpha$.

Thus only the first seminvariant λ_1 is changed; it becomes $\lambda_1 - a$, while λ_2, λ_3, ... are unaltered.

Change of Scale. If the scale of measurement is altered so that what was previously recorded as x now reads kx, then every factor t^x in the previous g.f. now becomes t^{kx}, that is, $(t^k)^x$. Hence in the m.g.f. the previous $e^{\alpha x}$ now reads $e^{\alpha kx}$. Hence the rules :

Change of scale, so that x *becomes* kx, *has the effect of replacing* t *by* t^{k} *in the probability g.f.,* α *by* $\mathrm{k}\alpha$ *in the m.g.f.*

The immediate consequence is that the previous r^{th} moment μ_r' and r^{th} seminvariant λ_r become $k^r\mu_r'$ and $k^r\lambda_r$.

The reason for the name *seminvariant* is now seen; for under a change of origin and scale in x all seminvariants after λ_1 are altered at most by a scale factor.

Change of scale in the f.m.g.f. will be effected by replacing $1+\alpha$ by $(1+\alpha)^k$.

Example. The first moment or mean of the distribution

which has g.f. $(pt+q)^n$ is np, and the m.g.f. with respect to the mean as origin is $e^{-npa}\ (pe^a+q)^n$. The corresponding f.m.g.f. is $(1+a)^{-np}\ (1+pa)^n$.

12. Population, Universal, Universe or Stock, Sample. To conclude these questions of nomenclature and general notions we explain what is meant by *population*, *universe* or *stock*, and *sample*. As an example let us consider the repetition of an experiment in which the probability of success is $p = m/n$, a rational fraction. We may construct a *model* by taking (or imagining) n similar objects, such as equal spherical marbles, of which m are distinguishable from the rest, and drawing an object repeatedly, with replacement after each drawing. Such an assemblage, actual or hypothetical, constitutes a *population*, *universe* or *stock*. It is in fact merely a model of the system S. To cope with special cases we have often to conceive a fictitious infinite population. For example, if we wish to represent drawing with replacement by a model in which the drawing is without replacement, the population of the model will certainly have to be infinite, since the probabilities of successive drawings are constant, a thing which cannot happen with a finite population.

Sample. Any element of a population is a *sample* of that population. For example, if five drawings are made, with replacement each time, from six cards numbered 1, 2, 3, 4, 5, 6, the population of possible sets of five cards contains 6^5 or 7776 elements, of which (3, 5, 5, 4, 1) and (4, 4, 2, 6, 3) are two samples. If the drawing is without replacement, the population of sets of five contains 6.5.4.3.2 or 720 elements, of which (2, 3, 5, 6, 4) and (5, 2, 4, 3, 1) are two samples. Or again, if a coin is spun 100 times, the sequence of heads and tails arising is to be regarded as one sample out of the possible 2^{100} sequences constituting the population of sequences.

The word " sample " is also used as a verb, " to sample "

a population meaning to draw a sample, or samples, from that population.

Notation. It is important to distinguish the probability $\phi(x)$, which may not be definitely known, from the relative frequency of x as found in a sample, let us say $f(x)$; and in the same way all parameters, such as means and moments, connected with $\phi(x)$ should be distinguished from the corresponding parameters in the case of $f(x)$. As far as possible we shall make this distinction by using Greek letters for probability functions and parameters, italic letters for the corresponding frequency functions and parameters. Thus if μ_r' stands for the r^{th} moment of $\phi(x)$, then m_r' will be the r^{th} moment of $f(x)$; and so on.

For detailed description of many aspects of theoretical and practical statistics, and for bibliographical references to memoirs and texts on the subject, the reader may consult *An Introduction to the Theory of Statistics*, by G. U. Yule and M. G. Kendall, London, 1937, the 11th edition of the original book by the first-named author.

CHAPTER II

PROBABILITY AND FREQUENCY DISTRIBUTIONS: GRAPHICAL REPRESENTATION: CALCULATION OF MOMENTS

13. Distributions, Probability Curve, Histogram. The assemblage of values of probabilities $\phi(x)$, for all the possible values x_j of x that may occur in any system S, is called the *probability distribution* of x in S. In practice a set of n observations in a sample does not usually give all the possible values x_j, and certainly cannot give them all if they cover a continuous range. Further, the sample of n values is itself only one member of the population, often prodigiously large or even infinite, of possible samples of n values that might have been drawn.

The relative frequency of x_j in a sample of n values is denoted by $f(x_j)$. The assemblage of relative frequencies $f(x_j)$ for the sample is then called the *frequency distribution* of x in that sample. The name is also often given to the assemblage of absolute or actual frequencies, but these are merely obtained by multiplying all relative frequencies by n.

Ex. 1. In repeated throws of a symmetrical coin the respective probabilities of runs of 1 head, 2 heads, 3 heads, ... are $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$, ... Hence in 400 throws the ideal probability distribution may be tabulated (to the nearest integer) as:

x	1	2	3	4	5	6	7	8	Total
$n\phi$	50	25	13	6	3	2	1	0	100

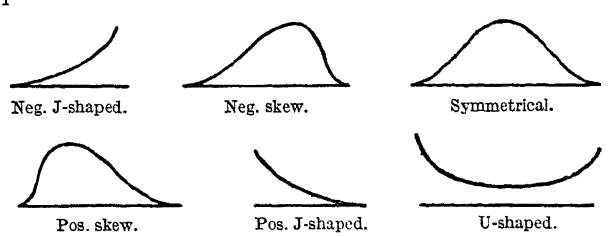
In an actual experiment of 400 throws (performed by the

author) there were 196 heads, and the frequency distribution distributions of runs of x heads was :

x	1	2	3	4	5	6	7	8	Total
nf	51	24	14	4	5	1	0	1	100

Comparing the actual with the theoretical distribution, the reader will note a fairly close agreement, and also a slight irregularity in the frequencies.

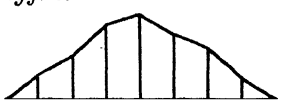
If x is a continuous variate, the curve $y = \phi(x)$ is called the *probability curve* of x. (The term " frequency curve " will often be found, but it is not strictly accurate. *Cf.* **12.**) The curve may be *symmetrical* about its central ordinate ; or it may have the " long tail " to the positive or right side, in which case it is said to be *positively skew* ; or to the negative or left side, in which case it is *negatively skew*. In some cases, as in the probabilities of runs of heads just considered, the curve may not descend at all on one side or the other. A curve so extremely skew is called *positively J-shaped*, or *negatively J-shaped*, as the case may be. In a rare type of distribution called the *U-shaped* curve the minimum ordinate is in the middle region. The *area* under a probability curve measures the total probability of all possible values of x, and is therefore equal to 1.



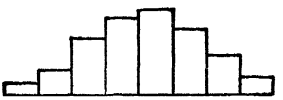
If x is a discontinuous variate the plotted points (x, y), where $y = \phi(x)$, do not form a curve. The sum of the ordinates is equal to 1. It is customary, though there is no very cogent reason for doing so, to join these points

each to its neighbour by straight lines, thus obtaining the *probability polygon* for the distribution in question. The terms "symmetry" and "skewness" then have corresponding meanings.

Frequency Polygon, Histogram. In an actual sample of observations we have relative frequencies instead of probabilities. If the variate x is discontinuous, as for example the number of flowers on stalks, the number of beans in bean-pods, we obtain separate plotted points $(x, f(x))$ which, joined to their neighbours, form a *frequency polygon.*



Frequency Polygon.



Histogram.

On the other hand, x may be a continuous variate, the range of which in the process of measurement is broken for convenience into intervals of finite breadth. For example, height of men, measured in inches, is a continuous variate ; all heights within a certain range are conceivable. But in practice heights may be recorded to the nearest inch, in which case all individuals of the sample having heights in the range 66·5000... to 67·4999... inches form a frequency group or *frequency class* corresponding to $x = 67$, the *central* point of the class. In such a case it is customary to represent the class graphically not by a single ordinate at the central point but by a rectangle on the class-interval (as 66·5 to 67·5) as base and of height proportional to the class frequency or relative frequency $f(x)$. The figure of juxtaposed rectangles is then called the *frequency histogram* or simply the *histogram* (that is, diagram made up of *cells*), and it furnishes a rough approximation to the ideal probability curve.

Ex. 2. Plot the probability polygon for the runs of heads in Ex. 1 ; also the frequency polygon of the experiment.

Ex. 3. Note that often great care must be taken to

ascertain the exact class-boundaries and centres of classes. For example, the British Anthropometric Committee (*Report*, 1883, p. 256) measured the height of 8585 adult males in the British Isles, made up of samples of 6194 from England, 1304 from Scotland, 741 from Wales and 346 from Ireland. The distribution of the Irish sample reads as follows:

x	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	n
nf	1	0	2	2	7	15	33	58	73	62	40	25	15	10	3	346

When we are told, however, that the class $x = 59$ inches means "59 and over," but at the same time that measurements were to the nearest eighth of an inch, it appears that class $x = 59$ means from $x = 58\frac{15}{16}$ to $59\frac{15}{16}$, so that the centre of the class is at $x = 59\frac{7}{16}$; and so for every other class.

The reader should draw the histogram for the above distribution, choosing not too small a scale for the frequency.

For ease and rapidity in computation we can always by a *change of origin* take any convenient value of x as new origin, and by a *change of scale* make class intervals of *unit breadth*. At the end of any calculations we can translate the results back to the proper origin and scale. It is often convenient to choose a *provisional origin* either near the middle values of x or at one or other end of the range.

Ex. 4. In the distribution of Ex. 3, if 67 is taken as new origin for x, the classes range from $x = -8$ to $x = +6$. If these classes are presumed to be centred, the origin is not 67 but $67\frac{7}{16}$.

14. Descriptive Parameters of Distribution. Probability and frequency distributions may be described, not completely, but in their main features, by the values of their moments, factorial moments or other parameters. Some of these parameters have a geometrical significance.

Typical Parameters or Averages. There are three of these in common use, the *mode*, the *median* and the *arithmetic mean*.

Mode. The mode is the value of x for which the probability $\phi(x)$, or in a frequency distribution the relative frequency $f(x)$, is a *maximum*, that is, greater than the probability (or frequency) on either side. In a probability curve it is the abscissa of a maximal ordinate.

Many curves have a single maximum near the middle; others may show two maxima or more. These are called *dimodal* or *multimodal*, as the case may be.

Mode.

Dimodal Curve.

Median.

Median. The median is that value of x which divides the sum or integral of the probabilities over the whole range into two equal parts. This sum or integral must be equal to 1; and so if the range of values of x is from $x = a$ to $x = b$, the median value of x is defined by

$$\sum_{a}^{x}\phi(x) = \sum_{x}^{b}\phi(x) = \tfrac{1}{2}\sum_{a}^{b}\phi(x) = \tfrac{1}{2}, \quad . \quad . \quad (1)$$

or

$$\int_{a}^{x}\phi(x)dx = \int_{x}^{b}\phi(x)dx = \tfrac{1}{2}\int_{a}^{b}\phi(x)dx = \tfrac{1}{2}. \quad . \quad (2)$$

For a continuous probability curve the median ordinate, by (2), *bisects the area* under the curve.

Arithmetic Mean. The most widely used typical measure is the arithmetic mean, which is simply the *first moment* or *mathematical expectation* of x, namely

$$\mu_1' = \Sigma x\phi(x) \text{ or } \int x\phi(x)dx. \quad . \quad . \quad (3)$$

These formulæ are the same as those occurring in dynamics for the *centroid* of a series of particles of masses $\phi(x_j)$ placed at points x_j along a straight line, and the centroid of a straight rod of density $\phi(x)$ at the point x. It follows that the arithmetic mean is the abscissa of the

ordinate through the centroid of the area under the curve $y = \phi(x)$.

The arithmetic mean of the values x_j in a sample is correspondingly $m_1' = \Sigma x f(x)$.

Remark. In many probability curves of slight or moderate skewness the median lies between the mode and the arithmetic mean, nearly twice as far from the mode as from the mean.

Moments about the Mean. The arithmetic mean is so fundamental in theory and in practice that it is customary, once it has been determined, to take it as a new origin and to refer all higher moments to this origin. Moments about the mean as origin are usually denoted by undashed μ_r. We find easily, by binomial expansion,

$$\begin{aligned}\mu_r &= \Sigma(x-\mu_1')^r\phi(x) \text{ or } \int (x-\mu_1')^r\phi(x)dx \\ &= \mu_r'-r\mu_1'\mu_{r-1}'+r_{(2)}(\mu_1')^2\mu_{r-2}'-\cdots \\ &\quad +(-)^{r-1}r(\mu_1')^{r-1}\mu_1'+(-)^r(\mu_1')^r, \quad . \quad (4)\end{aligned}$$

where $r_{(s)}$ denotes the familiar binomial coefficient $r(r-1)\ldots(r-s+1)/s!$. The last two terms can be merged into one as $(-)^{r-1}(r-1)(\mu_1')^r$. For example:

$$\begin{aligned}\mu_1 &= 0, \\ \mu_2 &= \mu_2'-(\mu_1')^2, \\ \mu_3 &= \mu_3'-3\mu_1'\mu_2'+2(\mu_1')^3, \\ \mu_4 &= \mu_4'-4\mu_1'\mu_3'+6(\mu_1')^2\mu_2'-3(\mu_1')^4, \quad . \quad (5)\end{aligned}$$

formulæ of regular application in practical work, since they hold equally well for moments of a frequency distribution, $\phi(x)$ being then replaced by $f(x)$, and μ by m.

Other means, such as the geometric and the harmonic mean, are very occasionally used with respect to rather special distributions.

Seminvariants in Terms of Moments about the Mean. By expanding the logarithm of

$$M(\alpha) = e^{\mu_1'\alpha}(1+\mu_2\alpha^2/2!+\mu_3\alpha^3/3!+\ldots)$$

as a series in powers of a, and comparing the coefficients of these with the coefficients in **10** (7), we find the relations between seminvariants (or cumulants) and moments about the mean. The first four relations are

$$\lambda_1 = \mu_1', \quad \lambda_2 = \mu_2, \quad \lambda_3 = \mu_3, \quad \lambda_4 = \mu_4 - 3\mu_2^2.$$

15. Measures of Dispersion or Spread. Distributions differ according as the values of x are spread densely or widely on either side of the mean. To describe this feature numerically we need parameters measuring *dispersion.*

The arithmetic mean of the deviations $x_j - \mu_1'$ from the mean is of course of no use for the purpose, being equal to zero. A measure occasionally used, but now falling into disuse, is the *mean absolute deviation* (the former name was " mean error ") defined by the arithmetic mean of deviations from the mean *all taken with positive sign,* namely

$$\Sigma|x-\mu_1'|\phi(x) \text{ or } \int |x-\mu_1'|\phi(x)dx, \quad . \quad . \quad (1)$$

where $|x-\mu_1'|$ denotes the positive numerical, or *absolute,* value of $x-\mu_1'$.

Though usually computed with respect to μ_1', it is actually in closer association with the median, in virtue of a certain *minimal* property, namely :

The median value of x *is such that the sum of the absolute deviations from it,* $\Sigma|x-x_j|$, *is a minimum.*

The median of a discrete set of values x_j needs more precise definition. If an *odd* number of values is ranged in monotonic order $x_0, x_1, \ldots, x_{2n}$ so that each $x_{j+1} \geqslant x_j$, we shall define the median as the middle value, x_n. If an *even* number of values is so arranged as $x_0, x_1, \ldots, x_{2n-1}$, we shall say that the median is *any* value of x in the middle interval, that is, is such that $x_{n-1} \leqslant x \leqslant x_n$. The minimal property may then be proved as follows :

(*a*) Let there be $2n+1$ values $x_0, x_1, \ldots, x_{2n}$, and let

us call the interval between x_{j-1} and x_j inclusive the j^{th} interval. The median is at $x = x_n$. Let us denote by $S(x)$ the sum $\Sigma|x-x_j|$ of absolute deviations from any x.

First consider $S(x)$ as compared with $S(x_n)$, where x is in the $(n+1)^{th}$ interval, on the right of the median, and $x-x_n = h$. Then the absolute deviations of the $n+1$ values $x_0, x_1, \ldots, x_n$ on the one side have each been increased by h, while those of the n values $x_{n+1}, x_{n+2}, \ldots, x_{2n}$ on the other have each been decreased by h. Hence in this interval

$$S(x)-S(x_n) = h. \qquad (2)$$

Now suppose x moves into the next interval, the $(n+2)^{th}$. Comparing $S(x)$ with $S(x_{n+1})$, we note that if $x-x_{n+1} = h$ the absolute deviations of the $n+2$ values $x_0, x_1, \ldots, x_{n+1}$ each receive an increment h, while those of the remaining $n-1$ values receive a decrement h. Hence in this interval

$$S(x)-S(x_{n+1}) = 3h. \qquad (3)$$

In this way $S(x)$ increases as x moves through successive intervals to the right, the increments which it receives within the intervals being $h, 3h, 5h, \ldots, (2n-1)h$; and by symmetry, or by a similar proof, $S(x)$ receives corresponding increments as x moves through successive intervals to the left of x_n.

Hence $S(x)$ is a minimum for $x = x_n$.

(*b*) Let there be $2n$ values $x_0, x_1, \ldots, x_{2n-1}$.

The reader will see at once that if x lies in the central interval, the n^{th} interval, and if within that interval is displaced by an amount h, then n absolute deviations on the one side each receive an increment h, while n on the other each receive a decrement h. Hence $S(x)$ is constant within the central interval.

Also, as x moves out of the central interval either to right or to left through successive intervals, $S(x)$ receives

the respective increments $2h$, $4h$, ..., $(2n-2)h$. Hence $S(x)$ is a minimum within the central interval.

(*c*) The result for a continuous variate x can be proved as a limiting case of (*a*) and (*b*), or else directly thus :

Let $(-a, b)$ be the range of values of x, the median being taken as the origin $x = 0$, so that

$$\int_{-a}^{0} \phi(x)dx = \int_{0}^{b} \phi(x)dx = \tfrac{1}{2}. \quad . \quad . \quad (4)$$

The integral $S(h)$ of absolute deviations from $x = h$, $h>0$, is then

$$S(h) = \int_{-a}^{h} (h-x)\phi(x)dx + \int_{h}^{b} (x-h)\phi(x)dx$$

$$= \left[\int_{-a}^{0} + \int_{0}^{h}\right](h-x)\phi(x)dx + \left[\int_{0}^{b} - \int_{0}^{h}\right](x-h)\phi(x)dx, \quad (5)$$

whereas

$$S(0) = \int_{-a}^{0} -(x)\phi(x)dx + \int_{0}^{b} x\phi(x)dx. \quad . \quad . \quad . \quad . \quad (6)$$

Hence

$$S(x)-S(0) = h\left[\int_{-a}^{0} \phi(x)dx - \int_{0}^{b} \phi(x)dx\right] + 2\int_{0}^{h} (h-x)\phi(x)dx$$

$$= 2\int_{0}^{h} (h-x)\phi(x)dx, \quad . \quad . \quad . \quad (7)$$

and this is essentially positive, since $\phi(x)$ is a positive function. The same result may be proved to hold for $h<0$, and so $S(0)$ is a minimum.

Note. The indeterminacy of the median of an *even* number of discrete values x_j matters exceedingly little in practice, the two middle values being for the most part indistinguishably close.

The Quartiles. The median ordinate halves the distribution. Halving again the two halves, we may find

values of x which are called the *quartile* measures. For discrete distributions, they lie one-quarter and three-quarters along the line of values x_j, supposed arranged in ascending order. For continuous distributions of range $x = a$ to $x = b$ they are values q_1, q_3 (it is hardly worth while here to press further Greek letters into service) such that

$$\int_a^{q_1} \phi(x)dx = \int_{q_3}^{b} \phi(x)dx = \tfrac{1}{4}. \qquad (8)$$

The median might be regarded as a middle quartile q_2, the other two are called the *upper* and *lower quartiles*. The value of $\frac{1}{2}(q_3-q_1)$ furnishes a measure of dispersion called the *semi-interquartile range*. Any value of x has an *à priori* probability of $\frac{1}{2}$ of being such that $q_1<x<q_3$; it is as likely to be inside the range as outside it. For this reason, in the theory of errors, this particular measure of dispersion has long been called the *probable error* of the distribution. The name is very misleading, since there is nothing specially probable about this particular deviation; and of late there has been a salutary tendency to supersede the so-called probable error by the standard deviation, which we now define.

Standard Deviation. The arithmetic mean of the *squared deviations* $(x-\mu_1')^2$ from the mean, that is, the second moment μ_2, is obviously a suitable measure of dispersion. The square root of this, $\sqrt{\mu_2}$, formerly called the *root-mean-square deviation*, is now called the *standard deviation* and is denoted by σ. The sample value is denoted by s. Thus $\sigma^2 = \mu_2$, $s^2 = m_2$.

Variance. Modern usage is tending more and more to treat μ_2 or σ^2 itself, rather than σ, as a suitable measure of dispersion, under the name of *variance*. We have therefore

$$\sigma^2 = \Sigma(x-\mu_1')^2\phi(x) \text{ or } \int (x-\mu_1')^2\phi(x)dx, \qquad (9)$$

while $$s^2 = \Sigma(x-m_1')^2 f(x) \quad . \quad . \quad . \quad . \qquad (10)$$

The standard deviation has also a *minimal* property, with respect to the arithmetic mean, namely :

The sum or mean of squared deviations is a minimum when taken with respect to the arithmetic mean.

This fact is obvious at once from the formula of **14** (5)

$$\mu_2 = \mu_2' - (\mu_1')^2$$

which shows that μ_2 can never exceed μ_2'.

Mean and Variance of Linear Function. If we distinguish the respective means and variances of three independent variates x, y, z by triple suffixes, thus, μ_{100}', μ_{010}', μ_{001}' and μ_{200}, μ_{020}, μ_{002}, then from the properties of seminvariants (**11**) the linear function $ax+by+cz$ has

$$\text{mean} \quad a\mu_{100}' + b\mu_{010}' + c\mu_{001}',$$

$$\text{variance} \quad a^2\mu_{200} + b^2\mu_{020} + c^2\mu_{002} ;$$

and similarly for a general linear function in any number of independent variates.

Range, Extremes. Other indications of the dispersion of a distribution are given by the size of the range of x itself, $b-a$, as well as by the highest value, b, or lowest value, a.

16. Measures of Asymmetry or Skewness. When the mean is taken as origin $x = 0$, it may happen that $\phi(x) = \phi(-x)$, so that the distribution is *symmetrical.*

Ex. 1. The distribution of number of heads in a throw of n symmetrical coins, described by the g.f. $(\frac{1}{2}t+\frac{1}{2})^n$, is symmetrical about $x = \frac{1}{2}n$.

Ex. 2. The continuous distribution described by

$$dp \quad \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-a)^2} dx$$

is symmetrical about $x = a$.

Ex. 3. The distribution given by

$$dp = \frac{1}{\pi} \frac{1}{1+x^2}$$

is symmetrical about $x = 0$.

Lack of symmetry, skewness, is revealed functionally or numerically in various ways.

Various Measures of Skewness. In a symmetrical distribution the distances of the quartiles q_1 and q_3 from the median q_2 will be equal. In a skew distribution the difference between these distances gives a coefficient of skewness, namely

$$\{(q_3-q_2)-(q_2-q_1)\}/\sigma = (q_3-2q_2+q_1)/\sigma,$$

the division by σ being for the purpose of removing arbitrary units of scale and obtaining an *absolute* coefficient.

A natural measure of skewness is however the third moment about the mean, μ_3. If the distribution is symmetrical $\mu_3 = 0$. If the long tail of the distribution is on the side of the positive values of x, the cubes of positive values of x outweigh the cubes of negative values, so that μ_3 is positive, and we have positive skewness. In the same way if the long tail of the curve is on the side of the negative values of x, then μ_3 is negative, and we have negative skewness.

To remove arbitrary units of measure, since μ_3 is of the dimensions of x^3, or of σ^3, we construct an *absolute* measure of skewness by dividing μ_3 by σ^3, that is by $\mu_2^{3/2}$. The square of this, μ_3^2/μ_2^3, is often denoted by β_1.

Another measure of skewness (due to K. Pearson) depends on the fact that in a skew curve the mean, median and mode are not the same. The measure in question is defined by

(Mean—Mode)/(Standard Deviation).

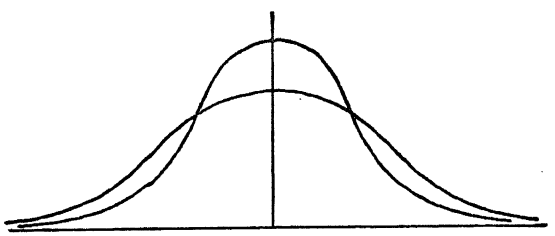
Like μ_3 it is positive for positive skewness, zero for symmetry, negative for negative skewness.

Skewness of Linear Function. If the 3rd moments of independent variates x, y, z about their means are μ_{300}, μ_{030}, μ_{003} respectively, the 3rd moment of $ax+by+cz$ about its mean is

$$a^3\mu_{300}+b^3\mu_{030}+c^3\mu_{003};$$

and similarly for linear functions of any number of independent variates. This follows (**14**) because $\mu_3 = \lambda_3$.

17. Measure of Flattening or Excess, Kurtosis. Two distributions may have the same mean, the same standard deviation, the same skewness, and yet may differ in that the curve of the one may be more flattened at the centre (*platykurtic*) than that of the other.



The degree of flattening is suitably measured by the 4th moment about the mean, μ_4. Removing arbitrary units of measure, just as in the case of β_1, we obtain the coefficient μ_4/μ_2^2, often denoted by β_2. It has been observed in an extensive class of probability curves, with scale chosen so that the variance is unity, that the ordinate at the mean or mode is greater or less according as β_2 itself is greater or less. Thus the value of β_2 serves to indicate whether the curve is tall and slim at the centre (*leptokurtic*) or squat (*platykurtic*). In the very important normal probability curve, which we shall meet in **32**, the value of β_2 is 3. Hence $\beta_2 - 3$ is sometimes called the *excess*, curves for which $\beta_2 < 3$ being *platykurtic*, those for which $\beta_2 > 3$ being *leptokurtic*, the normal curve being taken as standard.

Higher Moments. No simple geometrical interpretation attaches to parameters expressed by moments μ_r or m_r higher than the 4th, except of course that the moments of even order might be regarded as further measures of dispersion, and those of odd order as further

measures of skewness. These higher moments are in any case very seldom used in practice for *frequency* distributions, because being computed from values of x liable to random irregularity, " error " as it is usually called, they may be subject to very great error owing to the raising of some abnormally frequent large deviation x to a high power. This will be apparent when we come to consider the sampling error of coefficients, in Chapter VII.

18. Practical Computation of Moments. The initial stages of the analysis of frequency distributions almost always involve the computation of ordinary or factorial moments. In the case of a continuous variate artificially grouped (**13**) into classes, a certain error is introduced into the moments by the centring of class-frequencies about the centre of the class. The calculated moments then require adjustment by formulæ of rather wide application called *Sheppard's Corrections*.

The example on page 40 shows the computation of the first four moments and the coefficients of dispersion and excess, for a frequency distribution. The column headings explain themselves. It will be observed that transference is made to the more convenient *provisional* mean $x = 67$, this being judged by inspection of the distribution to be somewhere near the true mean.

Sheppard's corrections have not been used ; we shall allude to this example when we come to discuss them. As for the mean height of the group, the provisional origin is really, as we saw earlier, $67\frac{7}{16}$, or 67·44 inches. Hence the mean height is $67{\cdot}44+0{\cdot}34 = 67{\cdot}78$ inches.

The distribution shows a slight negative skewness. Whether this is a genuine effect or due to the irregularities of sampling cannot be decided until we know more about the probability distributions of coefficients calculated from samples.

The reader should verify that the sample estimates of β_1 and β_2 are ·0014 and 3·56.

Example. The distribution of heights of adult Irishmen.

X	nf	x	nfx	nfx^2	nfx^3	nfx^4
59	1	−8	−8	64	−512	4096
60	0	−7	0	0	0	0
61	2	−6	−12	72	−432	2592
62	2	−5	−10	50	−250	1250
63	7	−4	−28	112	−448	1792
64	15	−3	−45	135	−405	1215
65	33	−2	−66	132	−264	528
66	58	−1	−58	58	−58	58
→67	73	0	(−227) 0	0	(−2369) 0	0
68	62	1	62	62	62	62
69	40	2	80	160	320	640
70	25	3	75	225	675	2025
71	15	4	60	240	960	3840
72	10	5	50	250	1250	6250
73	3	6	18	108	648	3888
	346		(345)	)1668	(3915)	)28236
			)118	4·821	)1546	81·61
			0·341		4·468	

$m'_1 = 0{\cdot}341 + 67.$

$m_2 = 4{\cdot}821 - (0{\cdot}341)^2 = 4{\cdot}705.$

$m_3 = 4{\cdot}468 - 3(0{\cdot}341)(4{\cdot}821) + 2(0{\cdot}341)^3 = -0{\cdot}385.$

$m_4 = 81{\cdot}61 - 4(0{\cdot}341)(4{\cdot}468) + 6(0{\cdot}341)^2(4{\cdot}821) - 3(0{\cdot}341)^4 = 78{\cdot}84.$

19. Computation of Moments by Repeated Summation. If the origin of a distribution be taken at either end, preferably at the lower end, factorial moments can be computed by a process of repeated summation. We sum frequencies in columns from the remote value of x towards the origin, in the manner exemplified below. The leading sum in each column is one step lower than the leading sum in the preceding column.

Ex. 1. The same distribution, with origin at $x = 59$.

	x	nf	Σ	Σ^2	Σ^3	Σ^4	Σ^5
59 ←	0	1	346				
	1	0	345	2886			
	2	2	345	2541	11407		
	3	2	343	2196	8866	28343	
	4	7	341	1853	6670	19477	49757
	5	15	334	1512	4817	12807	30280
	6	33	319	1178	3305	7990	17473
	7	58	286	859	2127	4685	9483
	8	73	228	573	1268	2558	4798
	9	62	155	345	695	1290	2240
	10	40	93	190	350	595	950
	11	25	53	97	160	245	355
	12	15	28	44	63	85	110
	13	10	13	16	19	22	25
	14	3	3	3	3	3	3
		$r!$					24

The successive sums at the heads of columns may be proved (Appendix **2**) to be equal to $nm'_{(r)}/r!$. We have therefore

$$n = 346, \quad nm'_{(1)} = 2886, \quad nm'_{(2)} = 22814, \quad nm'_{(3)} = 170058,$$
$$nm'_{(4)} = 1194168.$$

Transforming to ordinary moments m'_r by the relations (Appendix **3**)

$$\begin{aligned} m'_1 &= m'_{(1)}, \\ m'_2 &= m'_{(2)} + m'_{(1)}, \\ m'_3 &= m'_{(3)} + 3m'_{(2)} + m'_{(1)}, \\ m'_4 &= m'_{(4)} + 6m'_{(3)} + 7m'_{(2)} + m'_{(1)}, \end{aligned} \qquad (1)$$

we obtain

$$\begin{aligned} m'_1 &= 2886/346 = 8{\cdot}34104, \\ m'_2 &= 25700/346 = 74{\cdot}2775, \\ m'_3 &= 241386/346 = 697{\cdot}647, \\ m'_4 &= 2377100/346 = 6870{\cdot}23, \end{aligned}$$

from which, by adjusting (**14** (5)) to moments about the mean, we derive

$$m_2 = 4{\cdot}705, \quad m_3 = -0{\cdot}386, \quad m_4 = 78{\cdot}87.$$

Now it may be noted that some advantage has been lost through the large numbers that arise in summations from end to end. Even though six significant digits have been retained throughout, the final results are very slightly discrepant with those computed by the other method. To obviate these disadvantages (which are not serious when a calculating machine is available) one may use either (i) factorial moments obtained by summation from both ends towards an origin near the centre, or (ii) central and mean central factorial moments obtained by a slight modification of this summation.

Ex. 2. Ordinary factorial moments. Origin at $x = 67$.

nf	Σ	Σ^2	Σ^3	Σ^4	Σ^5
1	1	1	1	1	1
0	1	2	3	4	5
2	3	5	8	12	17
2	5	10	18	30	47
7	12	22	40	70	117
15	27	49	89	159	276
33	60	109	198	357	633
58	*118*	*227*	*425*	*782*	*1415*
73	*228*				
62	155	*345*			
40	93	190	*350*		
25	53	97	160	*245*	
15	28	44	63	85	*110*
10	13	16	19	22	25
3	3	3	3	3	3
$r!$	1	1	2	6	24

From the italicized entries we obtain

$$n = 228+118 = 346,$$
$$nm'_{(1)} = 345-227 = 118,$$
$$nm'_{(2)} = (350+425)2 = 1550,$$
$$nm'_{(3)} = (245-782)6 = -3222,$$
$$nm'_{(4)} = (110+1415)24 = 36600.$$

These may be transformed to ordinary moments m'_r by the same relations as before, yielding

$$m'_1 = 118/346 = 0{\cdot}341, \qquad = 1668/346 = 4{\cdot}821,$$
$$m'_3 = 1546/346 = 4{\cdot}468, \qquad m'_4 = 28236/346 = 81{\cdot}61.$$

These are the same values as were found by the first method.

Ex. 3. Central and mean central factorial moments, with the same origin $x = 67$.

Here we again sum towards the centre from the ends, but each alternate sum (shown bracketed and italicized) involves the adding of only *half* the last summand in the preceding column, while the last sums in the other columns step successively away from the centre, as shown.

nf	Σ	Σ^2	Σ^3	Σ^4	Σ^5
1	1	1	1	1	1
0	1	2	3	4	5
2	3	5	8	12	17
2	5	10	18	30	47
7	12	22	40	70	117
15	27	49	89	159	276
33	60	109	198	*357*	(*454·5*)
58	118	*227*	(*311·5*)		
	(*154·5*)				
73					
	(*191·5*)				
62	155	*345*	(*522·5*)		
40	93	190	350	*595*	(*652·5*)
25	53	97	160	245	355
15	28	44	63	85	110
10	13	16	19	22	25
3	3	3	3	3	3
346	$r!$ 1	1	2	6	24

From the italicized entries we obtain the central factorial moments

$$n = 191{\cdot}5+154{\cdot}5 = 346,$$
$$nm_{\{1\}} = 345-227 = 118,$$
$$nm_{\{2\}} = (522{\cdot}5+311{\cdot}5)2 = 1668,$$
$$nm_{\{3\}} = (595-357)6 = 1428,$$
$$nm_{\{4\}} = (652{\cdot}5+454{\cdot}5)24 = 26568.$$

The formulæ for the m'_r in terms of the $m'_{\{r\}}$ are rather simple (Appendix **3**). We have

$$\begin{aligned} m'_1 &= m'_{\{1\}} \qquad 118/346 = 0{\cdot}341, \\ m'_2 &= m'_{\{2\}} = 1668/346 = 4{\cdot}821, \\ m'_3 &= m'_{\{3\}} + m'_{\{1\}} \quad (1428+118)/346 = 4{\cdot}468, \\ m'_4 &= m'_{\{4\}} + m'_{\{2\}} = (26568+1668)/346 = 81{\cdot}61, \end{aligned}$$

as before.

The moments about the mean can now be found in the usual way.

20. Sheppard's Corrections for Grouped Moments. As mentioned earlier, when a continuous distribution has been grouped into centred classes for convenience, the moments require adjustment or correction because of this artificial grouping. The necessary formulæ of correction were found by W. F. Sheppard.

Naturally the problem for perfectly general functions $\phi(x)$ is too broad, and it is necessary to impose conditions. Sheppard considered the case where $\phi(x)$ was such that the derivatives $\phi'(x)$, $\phi''(x)$, ... vanished in succession at the boundaries $x = a$ and $x = b$ to such an order that

$$\int_a^b x^r \phi^{(s)}(x)dx = w\sum_j x_j^r \phi^{(s)}(x_j) \qquad (1)$$

to a sufficient degree of accuracy, where w is the class-breadth and x_j the centre of a typical class; that is to say, the error committed should be negligible compared with sampling errors.

Remark. The relation between an *integral* and a *sum* of equidistant ordinates of the kind here considered enters into pure mathematics in the *Euler-Maclaurin summation formula*, by which a sum of ordinates is expressed as an integral over the range plus correction terms involving the derivatives of odd order taken at the boundaries. In many cases, where the derivatives $\phi'(x)$, $\phi'''(x)$, ... are not absolutely zero but converge to zero as a limit, the representation of the integral

on the left of (1) by the sum on the right needs very careful investigation. It is found, however, that for the statistical functions to which Sheppard's corrections are usually applied the difference between the integral and the sum can be made negligibly small by taking values of the class-interval w of a size quite customary in practice. Usually it is enough for w to be less than the standard deviation. The following derivation of the formulæ must be regarded as approximate only.

Ex. 1. The following two comparisons of integral with sum over an infinite range are interesting in this respect:

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi = 3{\cdot}14159 \text{ nearly},$$

whereas

$$\sum_{-\infty}^{\infty} \frac{1}{1+x^2} = 3{\cdot}15336 \text{ nearly},$$

x taking the values $0, \pm 1, \pm 2, \ldots$. Again,

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi} = 2{\cdot}506628275 \text{ nearly},$$

whereas

$$\sum e^{-\frac{1}{2}x^2} = 2{\cdot}506628288 \text{ nearly},$$

x taking the same values as before. The first sum in these examples is only moderately close to the corresponding integral; the second is very close, and still closer results are obtained if a summation with a finer subdivision of x is used.

Suppose the range $b-a$ divided into n class-intervals $(x-\frac{1}{2}w,\ x+\frac{1}{2}w)$, so that $b-a = nw$. If the probability in the j^{th} class is

$$p_j = \int_{x_j-\frac{1}{2}w}^{x_j+\frac{1}{2}w} \phi(x)dx, \qquad (2)$$

then the r^{th} moment calculated from the grouped classes is

$$\mu'_r = \sum_{j=1} x_j^r p_j, \qquad (3)$$

whereas the true moment is

$$\mu_r' = \int_a^b x^r \phi(x) dx. \tag{4}$$

Now

$$= \int_{x_j - \frac{1}{2}w}^{x_j + \frac{1}{2}w} \phi(x$$

$$= \int_{-\frac{1}{2}w}^{\cdot} \phi(x_j + x) dx$$

$$= \int_{-\frac{1}{2}w}^{\frac{1}{2}w} \{\phi(x_j) + x\phi'(x_j) + x^2\phi''(x_j)/2! + \ldots\} dx$$

$$= w\phi(x_j) + w^3\phi''(x_j)/24 + w^5\phi^{\text{IV}}(x_j)/1920 + \ldots, \tag{5}$$

provided that this series in powers of w converges.

Hence

$$\mu_r'' = \Sigma x_j^r p_j$$

$$= \int x^r \phi(x) dx + \frac{w^2}{24} \int x^r \phi''(x) dx + \frac{w^4}{1920} \int x^r \phi^{\text{IV}}(x) dx + \ldots, \tag{6}$$

in view of (1). Integrating by parts and using the fact that derivatives vanish at the boundaries, we have

$$\mu_r'' = \mu_r' + \frac{w^2}{24} r(r-1)\mu_{r-2}' + \frac{w^4}{1920} r(r-1)(r-2)(r-3)\mu_{r-4}' + \ldots \tag{7}$$

If moments μ_r about the mean are taken, we have therefore the relations (where μ_r'' means the r^{th} moment of the grouped classes about the mean) :

$$\begin{aligned}
\mu_0'' &= \mu_0 = 1, \\
\mu_1'' &= \mu_1 = 0, \\
\mu_2'' &= \mu_2 + \frac{1}{12} w^2 \mu_0 = \mu_2 + \frac{1}{12} w^2, \\
\mu_3'' &= \mu_3 + \frac{1}{4} w^2 \mu_1 = \mu_3, \\
\mu_4'' &= \mu_4 + \frac{1}{2} w^2 \mu_2 + \frac{1}{80} w^4, \ldots
\end{aligned} \tag{8}$$

which, on being solved for the μ_r, yield

$$\begin{aligned}
\mu_1 &= \mu_1' = 0, \\
\mu_2 &= \mu_2' - \frac{1}{12}w^2, \\
\mu_3 &= \mu_3', \\
\mu_4 &= \mu_4' - \frac{1}{2}w^2\mu_2' + \frac{7}{240}w^4,
\end{aligned} \qquad (9)$$

and these are the required adjustments, Sheppard's corrections. The correction to the second moment is especially simple and noteworthy. If the class-interval is taken as the unit of scale, the correction amounts to subtracting $\frac{1}{12}$ from the grouped second moment.

It is customary, though the practice requires more justification than it has ever received, to apply the same corrections for grouping in the case of frequency distributions, the presumption being that the moments thus corrected are a better representation of the moments of the underlying probability distribution.

Ex. 2. Correcting the moments about the mean for grouping in the example of **18** and **19**, we obtain for the corrected moments

$$\begin{aligned}
m_2 &= 4{\cdot}705 - 0{\cdot}083 = 4{\cdot}622. \\
m_3 &= -0{\cdot}385. \\
m_4 &= 78{\cdot}84 - \tfrac{1}{2}(4{\cdot}705) + 0{\cdot}029 \qquad 76{\cdot}52.
\end{aligned}$$

Ex. 3. The reader should seek out for himself numerous examples of frequency distributions, and should acquire as much practice as possible in computing moments in the various ways exemplified above, and in correcting them. Sheppard's correction will be applied in those cases in which the relative frequency $f(x)$ in sample corresponds to probability $\phi(x)$ of a *continuous* variate.

CHAPTER III

SPECIAL PROBABILITY DISTRIBUTIONS

21. Distributions of Equal Probability. If n values x_j of x, where $j = 1, 2, \ldots, n$, have each equal probability $1/n$, the graph of probability consists of n ordinates of equal height $1/n$. The case of a symmetrical coin is the case $n = 2$, the case of an ordinary unbiassed die is the case $n = 6$.

The Rectangular Distribution. The limiting case of the preceding, when n tends to infinity, yields an important distribution called the *rectangular distribution*, namely that in which x has an equal probability of being at any point in the range $x = a$ to $x = b$, $a<b$. The probability differential is then given by

$$dp = \frac{1}{b-a}\,dx, \tag{1}$$

so that $\phi(x) = 1/(b-a)$ and the probability curve consists of a rectangle on the range as base and of height $1/(b-a)$. It is always possible to choose the central point of the range for origin, and the unit of scale such that the range becomes the new range $x = -\frac{1}{2}$ to $x = \frac{1}{2}$. The rectangle is then a square. The moments of odd order vanish; those of even order are

$$\mu_r = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^r dx = \frac{1}{r+1}\left(\tfrac{1}{2}\right)^r. \tag{2}$$

In particular $\mu_2 = \frac{1}{12}$, so that $\sigma = 1/\sqrt{12} = 0{\cdot}2886\ldots$

Example 1. Show the m.g.f. of the standard rectangular distribution is $(\sinh \frac{1}{2}\alpha)/\frac{1}{2}\alpha$.

Example 2. The following samples from a rectangular population have been arranged as frequency distributions. The times on 1000 watches displayed in watchmakers' windows were noted by the author. The distributions are of the first and second 500 of these. Class $x = 1$ means the class of all watch times from 1 h. to 1 h. 59 m. to the nearest minute, and classes $x = 2, 3, \ldots 12$ have a similar meaning.

x	1	2	3	4	5	6	7	8	9	10	11	12	n
(i) nf	34	54	39	49	45	41	33	37	41	47	39	41	500
(ii) nf	47	41	47	49	45	32	37	40	41	37	48	36	500

The mathematical expectation of the number in any class is 500/12, or 42 to the nearest integer. One of the classes in the above samples contains 54, and another contains 32. We shall see later that the deviations here are not extreme.

The mathematical expectation, or mean of x in the population, is 6·5. The means of x in the above samples are 6·426 and 6·322.

22. The Binomial Distribution. This fundamental distribution arises when n trials are made of a constant system S with probability p of an event E, the number x of successes in the n trials being the variate. The g.f. is $(pt+q)^n$, and so by binomial expansion the probability function, namely the coefficient of t^x in the g.f., is

$$\phi(x) = n_{(x)}p^x q^{n-x}. \quad . \quad . \quad . \quad (1)$$

Moments. The f.m.g.f., obtained by putting $t = 1+a$, is seen at once to be $(1+pa)^n$, so that the mean, the coefficient of a in this, is np. Hence the f.m.g.f. about the mean is (11)

$$\begin{aligned}&(1+a)^{-np}(1+pa)^n\\ &= [(1+pa)(1-pa+p(p+1)a^2/2!-p(p+1)(p+2)a^3/3!+\ldots)]^n\\ &= [1+p(1-p)a^2/2!-2p(1-p)(p+1)a^3/3!+\ldots]^n, \quad . \quad (2)\end{aligned}$$

whence $\mu_{(2)} = npq$, $\mu_{(3)} = -2npq(p+1)$, so that

$$\mu_2 = \mu_{(2)} = npq, \ \mu_3 = \mu_{(3)}+3\mu_{(2)} = npq(q-p). \quad (3)$$

It is readily proved in the same way, by finding $\mu_{(4)}$ and hence μ_4, that

$$\mu_4 = npq[p^2+(3n-4)pq+q^2]. \quad . \quad . \quad (4)$$

The formula $\sigma = \sqrt{(npq)}$ is of fundamental importance.

Example. The following is a sample from a binomial population. The Swedish astronomer and statistician, C. V. L. Charlier, performed 1000 times the experiment of drawing 10 cards, one at a time with replacement after each drawing, from an ordinary pack, the number x of black cards in each set of 10 cards being the variate. Thus $n = 10$, $p = \frac{1}{2}$. He obtained the distribution

x	0	1	2	3	4	5	6	7	8	9	10	N
Nf	3	10	43	116	221	247	202	115	34	9	0	1000

The corresponding probability distribution has g.f. $(\frac{1}{2}+\frac{1}{2}t)^{10}$. Multiplying this by 1000 and recording the coefficients of powers of t to the nearest integer, we obtain

x	0	1	2	3	4	5	6	7	8	9	10	N
$N\phi$	1	10	44	117	205	246	205	117	44	10	1	1000

From Charlier's data we find $m_1' = 4{\cdot}933$, $m_2 = 2{\cdot}415$. The theoretical expectations are $\mu_1' = np = 5{\cdot}00$ and $\mu_2 = npq = 2{\cdot}5$.

We shall consider in a later section (**55**) whether these deviations of actual experimental results from theoretical expectation are reasonable under the hypothesis of random sampling.

23. The Binomial Distribution of Poisson. The ordinary binomial distribution is often called the Bernoullian distribution, after James Bernoulli, who first (in *Ars Conjectandi*, a work published in 1713, eight years after his death) investigated it in detail. S. D. Poisson in 1837 considered the problem of n trials, but with the system S varied each time so as to produce possibly different probabilities of success p_j, where $j = 1, 2, \ldots, n$. The g.f. is therefore

$$(p_1t+q_1)(p_2t+q_2) \ldots (p_nt+q_n), \quad . \quad . \quad (1)$$

and so the f.m.g.f. is

$$(1+p_1a)(1+p_2a) \ldots (1+p_na). \quad . \quad . \quad . \quad (2)$$

The coefficient of a in this f.m.g.f. gives us the mean, or mathematical expectation of the number of successes, as $p_1+p_2+\ldots+p_n$. Let us write

$$np = p_1+p_2+\ldots+p_n, \quad . \quad . \quad (3)$$

in order that we may later compare the moments with those of a Bernoullian distribution with the same mean probability p, and so characterized by the g.f. $(pt+q)^n$. The Poisson f.m.g.f. about the mean is (compare the details of **22** (2))

$$\begin{aligned} &\Pi_j(1+a)^{-p_j}(1+p_ja) \qquad (\Pi \equiv \text{product}) \\ &= \Pi[1+p_jq_ja^2/2!-2p_jq_j(p_j+1)a^3/3!+\ldots] \\ &= 1+\Sigma p_jq_ja^2/2!-2\Sigma p_jq_j(p_j+1)a^3/3!+\ldots \quad . \quad (4) \end{aligned}$$

Hence $\mu_{(2)} = \mu_2 = \Sigma p_jq_j$, and $\mu_{(3)} = -2\Sigma p_jq_j(p_j+1)$,

so that $\mu_3 = \mu_{(3)}+3\mu_{(2)} = \Sigma p_jq_j(q_j-p_j)$. . . (5)

24. Comparison of Bernoullian and Poissonian Variance. It will now be proved that the Poissonian variance, let us say σ_P^2, is less than the Bernoullian, σ_B^2. At first sight this may seem surprising, for one might imagine that the variation of probability of success in trials within the experiment would increase the variance of x, the number of successes. If we consider, however, the case of extreme variation of probability, namely the case in which some of the trials are certain of success, and the rest are certain of failure, we shall see that the smaller variance is natural enough; for in this extreme instance the value of x is constant and so its variance is zero.

The fact that the Poissonian variance is less than the Bernoullian is valuable, for it suggests a test for the constancy or otherwise of the system S from one trial

to the next, in other words, for statistical homogeneity within the experiment.

As in **23**, let p be the mean probability, $p = \Sigma p_j/n$. We have at once, by the usual transference to the mean,

$$\sigma_p^2 = \Sigma(p_j - p)^2/n = \Sigma p_j^2/n - p^2, \quad . \quad . \quad (1)$$

where σ_p^2 is the *variance of probability* in the n trials. Hence

$$\begin{aligned} \Sigma p_j q_j &= \Sigma p_j(1-p_j) \\ &= np - np^2 - \Sigma(p_j - p)^2 \\ &= npq - \Sigma(p_j - p)^2 \quad . \quad . \quad . \quad (2) \end{aligned}$$

that is,
$$\sigma_P^2 = \sigma_B^2 - n\sigma_p^2. \quad . \quad . \quad . \quad . \quad (3)$$

This result shows not only that the Poissonian variance is less than the Bernoullian, but by how much it is less.

25. The Lexian Distribution. The extension made by Poisson to the Bernoullian scheme consisted in varying the probability of success among the n trials, but *within* the experiment. A different kind of extension was considered by the German economist, W. Lexis, in 1877. The probability was taken by Lexis as constant in the n trials of one experiment, but as varying among k such experiments.

Let k Bernoullian sets of n repeated trials be made, each with constant probability of success within the set. Let p_i be the probability for the i^{th} set, where $i = 1, 2, \ldots, k$, and let x_i be the number of successes recorded in it. It is required to find the mean and variance of the distribution of the x_i.

The sets are here mutually exclusive, and the probability of each, if we imagine one of the p_i to be chosen and n trials to be then made, is $1/k$. Also the f.m.g.f. of x_i is $(1+p_i\alpha)^n$. Thus the f.m.g.f. of the Lexian distribution is

$$k^{-1}\sum_i(1+p_i\alpha)^n. \quad (1)$$

The coefficient of α shows that the mean is $n\Sigma p_i/k$. For

comparison with a repeated Bernoullian scheme let us put $np = n\Sigma p_i/k$. The f.m.g.f. about the mean is then

$$k^{-1}(1+\alpha)^{-np}\Sigma(1+p_i\alpha)^n \quad . \quad . \quad (2)$$
$$= [1-np\alpha+np(np+1)\alpha^2/2!+\ldots]$$
$$\times[1+np\alpha+\frac{n(n-1)}{k}\Sigma p_i^2\alpha^2/2!+\ldots]$$

whence, by picking out the coefficient of $\alpha^2/2!$,

$$\mu_2 = \mu_{(2)}$$
$$= np(np+1)-2n^2p^2+n(n-1)k^{-1}[kp^2+\Sigma(p_i-p)^2]$$
$$= npq+n(n-1)\Sigma(p_i-p)^2/k,$$

that is,

$$\sigma_L^2 = npq+n(n-1)\sigma_p^2. \quad . \quad . \quad . \quad (3)$$

Thus, whereas the Poissonian variance was less than the Bernoullian, we see that the Lexian variance *exceeds* the Bernoullian by an amount which increases strongly with n, because of the coefficient $n(n-1)$ in (3).

26. Coolidge's Extension of the Lexian Scheme. It is a natural extension to consider, as J. L. Coolidge did in 1921, the distribution which arises not from k Bernoullian but from k Poissonian sets, each with a different set of probabilities in its constituent n trials.

Let p_{ij} be the probability of success in the j^{th} trial of the i^{th} set. Then, just as in **25**, the f.m.g.f. is

$$k^{-1}\sum_i\prod_j(1+p_{ij}\alpha). \quad . \quad . \quad . \quad (1)$$

Let us write $\sum_j p_{ij} = np_{i0}$, $\sum_i p_{i0} = kp$. Then the mean of the distribution is evidently np. Transferring the f.m.g.f. to the mean, and picking out the coefficient of $\alpha^2/2!$, we find, after three or four lines of algebra,

$$\mu_2 = \mu_{(2)} = npq+n(n-1)\sum_i(p_{i0}-p)^2/k-\sum_{ij}\sum(p_{ij}-p_{i0})^2/k.$$

It is appropriate to regard the three terms of this expression as of Bernoullian, Lexian and Poissonian type respectively: Certain special cases are easily perceived; for example, when $p_{i0} = p$, that is to say, when the mean probability in each set of trials is the same for all sets, a variance emerges which slightly generalises the Poissonian variance σ_P^2, and, like it, is less than the Bernoullian.

An alternative form is

$$\mu_2 = npq + n^2 \sum_i (p_{i0} - p)^2/k - \sum\sum_{ij} (p_{ij} - p)^2/k,$$

which we may write as

$$\sigma_C^2 = \sigma_B^2 + n^2\sigma_{p_i}^2 - n\sigma_p^2. \quad . \quad . \quad . \quad (2)$$

This result shows that non-homogeneity, or fluctuation of probability, within the trials of an experiment is of far less effect, when n is large, than fluctuation in mean probability from one set to another. In fact in many cases σ_C^2 differs only slightly from the corresponding σ_L^2.

Analysis of Variance. The results which we have obtained for the Lexian and Coolidge schemes exhibit the variance as *resolved into separate components of variance.* The Bernoullian component may be called the *random* component, since it arises even when probability is constant, while the Lexian component may be called the *systematic* component, since it arises from the systematic alteration or variation of probability from one experiment to another. This resolution of variance into separate components of variance has been called *analysis of variance.* It has been greatly extended by Professor R. A. Fisher, who has devised regular schemes of experimental arrangement involving many variates, by means of which not one but several systematic components of variance can be isolated (75) from each other and from the random component.

27. Charlier's Criteria of Homogeneity Based on Dispersion. The test of homogeneity or stability considered in this section would now be superseded or

amplified by modern methods of analysis of variance, but it is interesting in itself.

We have approximately, in the Lexian and Coolidge schemes,

$$\sigma_L^2 = \sigma_B^2 + n^2\sigma_p^2. \quad . \quad . \quad . \quad (1)$$

Hence $$(\sigma_p/p)^2 = (\sigma_L^2 - \sigma_B^2)/(\mu_1')^2, \quad . \quad . \quad . \quad (2)$$

where $\mu_1' = np$, the mean of the distribution.

Hence $$\sigma_p/p = \sqrt{(\sigma_L^2 - \sigma_B^2)}/\mu_1'. \quad . \quad . \quad (3)$$

Charlier denoted this by ρ, naming it the " coefficient of perturbation " of a Lexian distribution. He turned it into a percentage by taking 100ρ. From (3) we see that ρ measures the *relative fluctuation* of probability.

Example. Classing 288,000 Swedish births in 576 sets of 500 each, according to different months and different districts, Charlier found for x, the number of male births in a set,

$$m_2' = 257{\cdot}12,\ s_L = 12{\cdot}49,\ n = 500,\ k = 576.$$

Hence $p = m_1/n = 0{\cdot}514$, $q = 0{\cdot}486$, not *à priori*, but as *estimated* from the large sample of 288,000 ; and so

$$s_B = \sqrt{(npq)} = \sqrt{124{\cdot}9} = 11{\cdot}18.$$

Hence $100\rho = 100(156{\cdot}0 - 124{\cdot}9)/257 = 2{\cdot}17$ per cent.

The conclusion made is that a male birth in Sweden is an event of 51·4 per cent. probability, with a standard deviation of $51{\cdot}4 \times 0{\cdot}0217$, or about 1·1 per cent. probability.

28. Types of Multinomial Distribution. The binomial distribution, of Bernoullian or Poissonian type, is a special case of the multinomial distribution, the forms of which are so many and so various as almost to defeat classification. We have seen a simple example in the probability distribution of totals of points in n throws of a die, or a single throw of n similar dice. Here the g.f., for biassed dice, is

$$(p_1t + p_2t^2 + p_3t^3 + p_4t^4 + p_5t^5 + p_6t^6)^n, \quad . \quad . \quad (1)$$

and it is best to leave the distribution in this symbolized

form, and not to expand by the multinomial theorem. The generalization to the case of n different dice, possibly with different numbers of faces, is easily seen.

Ex. 1. Prove that the mean value of the total in n throws of a biassed die is $n(p_1+2p_2+3p_3+4p_4+5p_5+6p_6)$.

Ex. 2. Find, by constructing the f.m.g.f., the variance and standard deviation of the total of points in a throw of n symmetrical six-sided dice.

The f.m.g.f. reduces to

$$\left[(1+\alpha)^{-5/2}\left(1+\frac{5}{2}\alpha+\frac{20}{3}\alpha^2/2!+\ldots\right)\right]^n$$

$$=\left[\left(1-\frac{5}{2}\alpha+\frac{35}{4}\alpha^2/2!+\ldots\right)\left(1+\frac{5}{2}\alpha+\frac{20}{3}\alpha^2/2!+\ldots\right)\right]^n$$

$$=\left(1+\frac{35}{12}\alpha^2/2!+\ldots\right)^n.$$

Hence $\mu_2 = 35n/12$, and so $\sigma = \sqrt{(35n)}/2\sqrt{3}$.

29. Sampling without Replacement, Hypergeometric Distribution. When in sampling a population the individual drawn is not replaced, the result of one drawing influences the probability of the next, so that the successive drawings are not independent. Hence it is no longer possible to combine into a product the g.f.'s of the separate drawings. It is true that the difficulty can be circumvented by the introduction of symbolic products, with due precautions in expansion, but we shall here proceed from first principles.

Let us consider a population of N individuals, of whom $M = Np$ are of character A, so that the probability of drawing an A at the first drawing is p. Let n drawings be made, no individual drawn being replaced after the drawing. It is required to find the probability distribution of x, the number of individuals A drawn.

The probability of x successes A, $n-x$ failures $\bar{A}$, occurring in some particular order, is

$$M(M-1)\ldots(M-x+1)(N-M)(N-M-1)\ldots$$
$$(N-M-n+x+1)/N(N-1)\ldots(N-n+1), \quad . \quad (1)$$

as is readily seen by considering how the numbers in population, and in categories A or $\bar{A}$, are depleted by 1 at each drawing. But there are $n_{(x)}$ possible orders in which x successes may eventuate among n drawings. Hence the desired probability is

$$\phi(x) = n_{(x)}M^{(x)}(N-M)^{(n-x)}/N^{(n)}, \quad . \quad (2)$$

where $M^{(x)} = M(M-1)(M-2) \ldots (M-x+1)$, and so on.

Just as the binomial probability function of 22 was a typical term in the binomial expansion of $(pt+q)^n$, so this function that we have just found is a typical term in a certain series, a *hypergeometric* series. (The series is however of higher type than the ordinary Gaussian hypergeometric series.) Hence $\phi(x)$ is often called the *hypergeometric* probability function.

The g.f. is

$$\sum_{x=0}^{n} n_{(x)}M^{(x)}(N-M)^{(n-x)}t^x/N^{(n)} \quad (3)$$

and so the f.m.g.f. is

$$\sum_{x=0}^{n} n_{(x)}M^{(x)}(N-M)^{(n-x)}(1+\alpha)^x/N^{(n)}, \quad (4)$$

which may be evaluated (with some trouble if only elementary methods are used) as

$$1+\frac{Mn}{N}\alpha+\frac{M^{(2)}n^{(2)}}{N^{(2)}}\alpha^2/2!+\frac{M^{(3)}n^{(3)}}{N^{(3)}}\alpha^3/3!+\ldots, \quad . \quad (5)$$

a terminating hypergeometric series which in the notation of Gauss would be written $F(-M, -n; -N; \alpha)$. The mean is thus Mn/N, and the r^{th} factorial moment is $M^{(r)}n^{(r)}/N^{(r)}$.

The examples which have now been given of probability distributions have shown how numerous and varied are the types of distribution. In fact, any proposed probability function may be simulated by a suitably constructed model or population, and special samplings of this population

give rise to further probability functions. Fortunately, when the number n of trials is large, many of these probability distributions tend with good approximation towards one or other of a few dominant types, which we shall now consider.

30. Important Approximate Distributions: Types A and B. When the coefficients of t^x in the Bernoullian binomial g.f. $(pt+q)^n$ are taken as probability ordinates $y = \phi(x)$, we may join the tops of the ordinates to form a probability polygon. If this is done for increasing values of n, the mean np being taken as origin and the standard deviation $\sqrt{(npq)}$ as unit of scale, it is found that the successive probability polygons tend to lose any initial asymmetry due to inequality of p and q.

In fact the coefficient β_1 of skewness is

$$\beta_1 = \mu_3^2/\mu_2^3 = [npq(q-p)]^2/(npq)^3 = (q-p)^2/npq, \quad . \quad . \quad . \quad (1)$$

which evidently tends to zero as n increases, unless either of p or q is of the order of magnitude of $1/n$, let us say $O(1/n)$, in which case the skewness remains appreciable. Not only so but, apart from the exception just mentioned, these binomial curves are found to cluster towards a *limiting symmetrical* shape, the same for all. The curve to which they thus approach asymptotically is of paramount importance in statistics, and is called the *normal* probability curve. It is the asymptotic shape not merely of the Bernoullian binomial but of the Poissonian, as well as of the multinomial and of many other distributions, and it is characterized by the probability differential

$$dp = \phi(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} dx, \quad . \quad . \quad (2)$$

where μ is the mean, σ the standard deviation.

When small corrective terms involving n are retained,

a closer representation is given by the probability function of Type A, namely

$$p(x) = \phi(x) - a_3\phi'''(x)/3! + a_4\phi^{iv}(x)/4! - \ldots, \qquad (3)$$

where $\phi(x)$ denotes the normal probability function, and the coefficients a_r of the derivatives $\phi^{(r)}(x)/r!$ are of irregularly decreasing orders of magnitude with respect to n. The coefficient a_3 measures skewness, a_4 measures excess.

As noted above, the case when p is very small is exceptional. If p is $O(1/n)$, the mean np is not $O(n)$ but $O(1)$. In this case the normal function is not the most suitable basis of approximation, and the appropriate asymptotic probability function is Poisson's function of statistical rareness, namely

$$\psi(x) = e^{-\mu}\mu^x/x!, \qquad (4)$$

where μ is the mean. Here again, when terms of smaller order involving n are retained, a closer representation is given by the probability function of Type B, namely

$$p(x) = \psi(x) + b_2\nabla^2\psi(x)/2! - b_3\nabla^3\psi(x)/3! + \ldots, \qquad (5)$$

where $\psi(x)$ is Poisson's function (4) above, and ∇ denotes the operation of forming the *receding difference*, so that $\nabla\psi(x) = \psi(x) - \psi(x-1)$. It proves to be the case that b_2 is $O(n^{-1})$, b_3 and b_4 are $O(n^{-2})$, b_5 and b_6 are $O(n^{-3})$, and so on.

We now consider the derivation of these functions.

31. The Normal Function as Limit of the Binomial. The rigorous derivation of the normal function as generated by compounding n independent distributions, and the discussion of necessary and sufficient conditions, require advanced mathematics beyond our scope. We content ourselves here with elementary and incomplete treatments.

Consider first the binomial g.f. $(pt+q)^n$, where p is

not of order $1/n$, but is $O(1)$. Putting $t = e^{\alpha}$ we have the m.g.f.

$$(pe^{\alpha}+q)^n = (1+p\alpha+p\alpha^2/2!+p\alpha^3/3!+)^n.... \quad . \quad (1)$$

The mean is np. Let us transfer to the mean, and to discover the limiting shape of the curve of probability let us alter the scale, so as to find the distribution, not of actual number of successes x, but of the deviation $(x-np)/n$ of the *relative* frequency of successes from the mean p of relative frequency.

As a first step we construct the m.g.f. of x/n. By **11** it is

$$[1+p\alpha/n+p\alpha^2/2!n^2+O(n^{-3})]^n$$
$$= [(1+p\alpha/n+\tfrac{1}{2}p^2\alpha^2/n^2)(1+\tfrac{1}{2}(p-p^2)\alpha^2/n^2+O(n^{-3}))]^n, \quad (2)$$

where $O(n^{-3})$ indicates in both cases remainder terms of order n^{-3}. As n increases this m.g.f. tends asymptotically to

$$e^{p\alpha}e^{\frac{1}{2}pq\alpha^2/n}. \quad . \quad . \quad . \quad . \quad (3)$$

The first factor shows that the mean of the transformed variate is p; but this we already know. The second factor indicates, by a further obvious transformation of scale, that the m.g.f. of the standardized deviation $z = (x-np)/\sqrt{(npq)}$ is

$$e^{\frac{1}{2}\alpha} \quad (4)$$

Now the possible number x of successes may range from 0 to n. Thus the values of z may range from $-\sqrt{(np/q)}$ to $+\sqrt{(nq/p)}$, a range which tends in both directions to infinity. Further, consecutive values of x differ by 1, and so consecutive values of z differ by $1/\sqrt{(npq)}$, an interval which tends to zero as n increases. We therefore seek a representation of the probability function $\phi(z)$ as a positive function *continuous* over the range $-\infty$ to ∞; and the question is, what function $\phi(z)$ is such that its m.g.f.

$$\int_{-\infty}^{\infty} \phi(z)e^{\alpha z}dz = e^{\frac{1}{2}\alpha^2} \text{ ?} \quad (5)$$

The answer is contained in a theorem, to the effect that the only positive continuous function satisfying this relation for some continuous range of values of α is

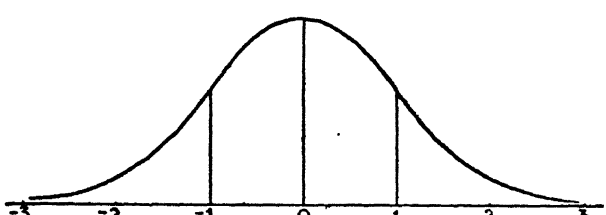
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \tag{6}$$

and this is the normal probability function, in standard form.

The reader should become thoroughly familiar both with this form and with the unstandardized form of **30** (2).

Incidentally, taking the logarithm of the m.g.f. (4), we see that apart from the mean or first seminvariant there is only one other seminvariant, namely λ_2 or σ^2.

32. Properties of the Normal Probability Function. The curve of the normal function is a symmetrical bell-shaped curve, extending to infinity on either side and flattening rapidly upon the axis of x.



The maximum ordinate is $y_0 = 1/\sqrt{(2\pi)}$. The area under the curve is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = 1, \tag{1}$$

by the well-known integral. (Gillespie, *Integration*, p. 88.) The points of inflexion, given by $d^2y/dz^2 = 0$, will be found to be at $z = \pm 1$, or, in unstandardized units, at deviations $\pm\sigma$ from the centre.

The probability, as taken from the normal curve, that

a deviation from the mean is numerically less than z is the area under the curve between the ordinates for $-z$ and $+z$, namely

$$\frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{1}{2}z^2} dz. \tag{2}$$

This function, called the *error function* or *probability integral*, is denoted by $\mathrm{erf}(z)$ and has been extensively tabulated. (It is called the error function because the typical distribution of errors committed by instruments of observation has been found to be sensibly normal.) The following short table shows how the probability of deviations outside the range $(-z, z)$ diminishes as z increases :

z	0	0·5	1·0	1·5	2·0	2·5	3·0	3·5	4·0
erf(z)	0	0·383	0·683	0·866	0·954	0·988	0·997	0·9995	0·99994

We may note that the probability of a deviation greater than σ is about 1/3 or more nearly 7/22 ; that of one greater than 2σ is about 1/20 or more nearly 1/22 ; that of one greater than 3σ is about 1/370 ; and that of one greater than 4σ is about 1/17000.

The quartile deviation or so-called " probable error " is given by

$$\frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{1}{2}z^2} dz = \tfrac{1}{2}. \tag{3}$$

By interpolation it is found to be $z = 0{\cdot}6745$ nearly, corresponding to a deviation from the mean of about 2/3 or more nearly 27/40 of the standard deviation.

The mean absolute deviation is given by

$$\frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} z e^{-\frac{1}{2}z^2} dz = \sqrt{(2/\pi)} = 0{\cdot}7979 \text{ nearly}, \tag{4}$$

corresponding to about 4/5 of the standard deviation.

The higher moments of the normal function are found by expanding the m.g.f. $\exp(\frac{1}{2}a^2)$, or the unstandardized $\exp(\frac{1}{2}\sigma^2 a^2)$, and observing the coefficients of $a^r/r!$. For

odd orders they vanish, for even orders $2r$ they are given by

$$\mu_{2r} = (\tfrac{1}{2})^r \sigma^{2r}(2r)!/r! \quad . \quad . \quad . \quad (5)$$

In particular

$$\mu_4 = 3\sigma^4,$$

so that (17) the coefficient of excess $\beta_2 = \mu_4/\mu_2^2 = 3$.

33. Poissonian Function of Rare Statistical Frequency. We return to the binomial g.f. $(pt+q)^n$, examining the previously excepted case in which, though n becomes large, p is so small that the mean np is $O(1)$; in fact $p = O(n^{-1})$. Writing the mean np as μ, we have $p = \mu/n$. The f.m.g.f. is therefore (22)

$$(1+\mu a/n)^n, \text{ which tends to } e^{\mu a} \quad . \quad . \quad (1)$$

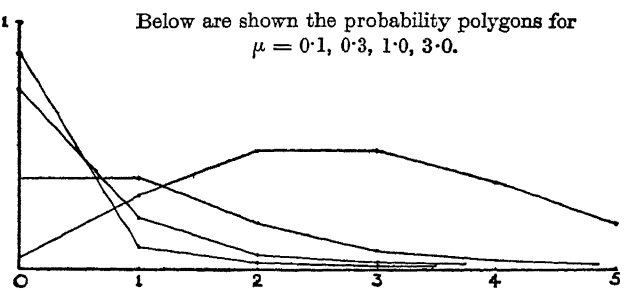
as n increases. This is the f.m.g.f. of the Poissonian function. The probability g.f. is therefore

$$e^{\mu(t-1)}, \quad . \quad . \quad . \quad . \quad (2)$$

and the coefficient of t^x in this gives the desired probability function as

$$\psi(x) = e^{-\mu}\mu^x/x!. \quad . \quad . \quad . \quad (3)$$

34. Properties of the Poissonian Function. The normal function contains two parameters, the mean μ and the standard deviation σ. The Poissonian function has one parameter only, the mean μ. The range of the



function is from $x = 0$ to $x = \infty$. For $\mu<1$ the probability polygon is J-shaped, for $\mu\geqslant 1$ it becomes double-sided and for large values of μ tends to acquire symmetry. Indeed, for large values of μ the shape is approximately normal; for the ordinary m.g.f. is

$$\exp[\mu(e^{\alpha}-1)] = \exp(\mu\alpha+\mu\alpha^2/2!+\mu\alpha^3/3!+\ldots) \qquad (1)$$

and if we change the scale so as to make $\sqrt{\mu}$ the unit we obtain the g.f.

$$\exp(\mu^{\frac{1}{2}}\alpha+\alpha^2/2!+\alpha^3/3!\mu^{\frac{1}{2}}+\ldots), \qquad (2)$$

which, to a first approximation (that is, including the first two terms of the series in the bracket) is the m.g.f. of a normal function with mean $\sqrt{\mu}$ and unit standard deviation.

The logarithm of (1) gives the seminvariant g.f. of the Poissonian function as $\mu(e^{\alpha}-1)$, which shows that all the seminvariants λ_r are equal to the mean μ; in particular the variance λ_2 or μ_2 is equal to μ.

There is only one factorial seminvariant, $\lambda_{(1)} = \mu$.

35. More General Derivations; Types A and B. As before, the extensions of the domain of application of the fundamental distributions given below are not established under the widest conditions.

Let us consider the compounding of n systems S_j, where $j = 1, 2, \ldots, n$, where each system has finite seminvariants, all $O(1)$. The seminvariant g.f. of S_j is then a convergent series

$$L_j(\alpha) = \lambda_1^{(j}\alpha+\lambda_2^{(j}\alpha^2/2!+\lambda_3^{(j}\alpha^3/3!+\ldots \qquad (1)$$

For example, the binomial distribution of n throws of a coin, provided that p is $O(1)$ and not $O(n^{-1})$, may be proved to have an s.g.f. of this kind.

Now imagine all the n systems S_j to operate independently, the results being added to make a variate x. By

the additive property of seminvariants the seminvariant g.f. of x is

$$\sum_j(\lambda_1^j\alpha+\lambda_2^j\alpha^2/2!+\lambda_3^j\alpha^3/3!+\ldots), \quad . \quad . \quad (2)$$

and the s.g.f. about the mean of x is the same with the term in α removed. The second seminvariant of x is clearly $O(n)$, and so the standard deviation is $O(n^{\frac{1}{2}})$. Let us therefore alter the scale so that $x/\sqrt{n}$ becomes the variate. The s.g.f. of this variate is then

$$\lambda_1\alpha+\lambda_2\alpha^2/2!+\lambda_3\alpha^3/3!+\ldots, \quad . \quad . \quad (3)$$

where λ_1 is $O(n^{\frac{1}{2}})$, λ_2 is $O(1)$, λ_3 is $O(n^{-\frac{1}{2}})$ and in general λ_r is $O(n^{1-\frac{1}{2}r})$. Again, the s.g.f. about the mean is the same with the term in α removed.

Thus as n increases the dominant term in the s.g.f. about the mean is $\lambda_2\alpha^2/2!$, which is the s.g.f. of a normal function

$$\phi(z) = \frac{e^{-\frac{1}{2}z^2/\lambda_2}}{\sqrt{(2\pi\lambda_2)}} \quad (4)$$

If, however, we retain the terms of smaller order, while choosing the scale so that $\lambda_2 = 1$, the m.g.f. about the mean is

$$\begin{aligned} M(\alpha) &= \exp(\tfrac{1}{2}\alpha^2)\exp(\lambda_3\lambda_2^{-\frac{3}{2}}\alpha^3/3!+\lambda_4\lambda_2^{-2}\alpha^4/4!+\ldots) . \\ &= \exp(\tfrac{1}{2}\alpha^2)(1+a_3\alpha^3/3!+a_4\alpha^4/4!+\ldots) \quad . \quad . \quad (5) \end{aligned}$$

where the second factor in brackets on the right arises from the expansion of the second exponential in the first line. Now if a probability function $P(x)$, which vanishes with all its derivatives at the boundaries, has m.g.f.

$$M(\alpha) = \int_a^b P(x)e^{\alpha x}dx \quad . \quad . \quad . \quad (6)$$

it may be proved by r integrations by parts that

$$\alpha^r M(\alpha) = (-)^r \int_a^b \left(\frac{d}{dx}\right)^r P(x)e^{\alpha x}dx. \quad . \quad . \quad (7)$$

Thus here, reverting the m.g.f. of (5) term by term, we derive the corresponding probability function as

$$p(z) = \phi(z) - a_3\phi'''(z)/3! + a_4\phi^{iv}(z)/4! - \ldots, \qquad (8)$$

provided that the series for m.g.f. and probability function are convergent. This is the probability function of Type A.

A close examination of the magnitude of terms in the expansion of

$$\exp(\lambda_3\lambda_2^{-\frac{3}{2}}\alpha^3/3! + \lambda_4\lambda_2^{-2}\alpha^4/4! + \ldots) \quad . \quad . \quad (9)$$

shows that the order of magnitude of coefficients in the series of Type A is as follows:

$$a_3 = O(n^{-\frac{1}{2}}),\ a_4 \text{ and } a_6 = O(n^{-1}),\ a_5,\ a_7 \text{ and } a_9 = O(n^{-\frac{3}{2}}),$$

and later coefficients show a similar irregularity.

Here let us pause to point out a practical disadvantage of the representation by Type A. If we are representing a given frequency distribution by Type A, we must use the *observed* moments to *estimate* the coefficients a_3, a_4, ... in Type A. Let us suppose that the convergence demands the retention of terms up to $O(n^{-1})$. We must then include not only a_4 but also a_6. Now a_6 depends on the 6th moment, and the 6th moment of the observations is subject to very high sampling error (**68**). Hence the effort to increase *mathematical accuracy* by retention of higher terms is largely frustrated by the *statistical inaccuracy* of the moments used to estimate those terms.

Series of Type B. The procedure for deriving the function of Type B is rather similar. The f.m.g.f. proves to be

$$\exp(\mu\alpha)(1 + b_2\alpha^2/2! + b_3\alpha^3/3! + \ldots), \quad . \quad . \quad (10)$$

which on reversion term by term gives

$$p(x) = \psi(x) + b_2\nabla^2\psi(x)/2! - b_3\nabla^3\psi(x)/3! + \ldots, \qquad (11)$$

the series of Type B, where $\psi(x)$ denotes Poisson's function

of **33** (3). Here the order of magnitude of coefficients is found to be :

$$b_2 = O(n^{-1}),\ b_3 \text{ and } b_4 = O(n^{-2}),\ b_5 \text{ and } b_6 = O(n^{-3}),$$

and so on. Thus in using the function of Type B for the representation of a frequency distribution it is best to truncate the series after a difference of *even* order.

36. Other Systems of Probability Functions : the System of Pearson. We have seen how the functions of Types A and B arise by the addition of seminvariant (or factorial seminvariant) generating functions, corresponding to the compounding of values of an additive variate. But a variate of this kind is a very special one. For example, if x is built up of added increments, then x^2, which we might have occasion to use instead of x, is certainly not the sum of the squares of those increments. Indeed, as we may well anticipate, the distribution of x^2 is different from that of x.

For this and for other reasons the scope of typical probability functions has been widened, and systems other than Type A and Type B have found acceptance. One such system is the system introduced in 1895 by Karl Pearson.

Let us consider the difference or differential equations satisfied by some of the standard probability functions. We shall use the receding difference operation defined by $\nabla\phi(x) = \phi(x) - \phi(x-1)$.

(i) The binomial probability function of **22** (1) satisfies

$$\nabla\phi(x) = -\frac{x-(n+1)p}{p(n-x+1)}\,\phi(x). \quad . \quad . \quad (1)$$

(ii) The Poissonian function $\psi(x)$ of **30** (4) satisfies

$$\nabla\psi(x) = -\frac{x-\mu}{\mu}\,\psi(x). \quad . \quad . \quad (2)$$

(iii) The hypergeometric probability function of **29** (2)

$$\nabla\phi(x) = -\frac{x(N+2)-(M+1)(n+1)}{(M-x+1)(n-x+1)}\phi(x). \qquad (3)$$

(iv) The normal probability function in standard form **31** (6) satisfies

$$\frac{d}{dx}\phi(x) = -x\phi(x). \quad . \quad . \quad . \quad (4)$$

A number of other probability functions, arising naturally in problems of repeated trials, might be added to this list. The Pearsonian system consists of the functions $\phi(x)$ which satisfy the differential equation

$$\frac{dy}{dx} \qquad \frac{(x-a)y}{c_0+c_1x+c_2x^2} \quad . \quad (5)$$

The functions are found by immediate integration; thus

$$\log y = -\int \frac{(x-a)dx}{c_0+c_1x+c_2x^2}, \quad . \quad . \quad (6)$$

whence y can be found by the methods of elementary integral calculus. The quadratic in the denominator of the integrand may have real, variously positive or negative, or equal, or numerically equal but of opposite sign, or complex roots; or again, with $c_2 = 0$, may degenerate into a linear function, or with c_1 and $c_2 = 0$ into a constant. These various cases yield the Pearsonian curves, usually classified into twelve types; while the discriminant of the quadratic, expressed in terms of moments of the curves, yields a "criterion" for judging in advance what type is appropriate to a proposed frequency distribution.

A full account of the curves, their shape and the process of representing frequency data by them is given in Elderton's *Frequency Curves and Correlation* (3rd edition, London, 1938), to which we refer the reader for details.

Here we have space to mention from time to time only a few of the curves, as they occur in special problems.

37. Probability Functions Generated by Change of Variate. If x is distributed about the mean $x = 0$ in a normal distribution

$$dp = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} dx, \quad . \quad . \quad . \quad (1)$$

it is certainly not the case that x^2 is normally distributed; for putting $z = \frac{1}{2}x^2$, we have $dx = (2z)^{-\frac{1}{2}} dz$, and so

$$dp = \pi^{-\frac{1}{2}} z^{-\frac{1}{2}} e^{-z} dz. \quad . \quad . \quad . \quad (2)$$

The range of z is from 0 to ∞, and the constant $\pi^{-\frac{1}{2}}$ is such that the integral of the probability function of z over this range is 1. The distribution of z is skew, and is actually a case of Pearson's Type III.

Ex. 1. Prove that the m.g.f. of z is $(1-\alpha)^{-\frac{1}{2}}$.

Again, if x is distributed between $-\frac{1}{2}$ and $\frac{1}{2}$ in the rectangular distribution $dp = dx$, the cube root $z = x^{\frac{1}{3}}$ is distributed, as the reader should verify, in the U-shaped distribution $dp = 3z^2 dz$. Or again, to take an example from physics, if the distribution of the velocities of a great number of particles about a zero mean velocity were normal, the distribution of their energies would be of Type III.

The derivation of probability functions from the normal function by non-linear change of variate was emphasized by J. C. Kapteyn in 1903 (*Skew Curves in Biology and Statistics*, Groningen), but was by no means a new conception even at that time.

Ex. 2. If x is a normal variate in standard measure, we have seen in Ex. 1 that the m.g.f. of $z = x^2$ is $(1-\alpha)^{-\frac{1}{2}}$. Hence the m.g.f. of $x_1^2 + x_2^2 + \ldots + x_n^2$, where the x_i are independent normal variates with the same mean $x = 0$ and in standard measure, is $(1-\alpha)^{-\frac{1}{2}n}$. The probability function which has this m.g.f. is unique, and of the form $cz^{\frac{1}{2}(n-2)} e^{-\frac{1}{2}nz}$.

The reader should verify that this function actually has the above m.g.f., and should find by integration the value of c.

Ex. 3. If x is distributed normally about $x = 0$ as mean, find the distributions of: (i) $z = e^x$, (ii) $z = x^2$, (iii) $z = x^{\frac{1}{2}}$.

38. Cauchy's Probability Function. The probability function which we shall next consider arises by change of variate in a rectangular distribution. Let us take a point Q on the axis of y at unit distance from the origin O. Let a straight line be taken at angle θ to QO, all values of θ from $-\frac{1}{2}\pi$ to $\frac{1}{2}\pi$ being equally likely, to cut the x axis in the point $X = (x, 0)$. What is the probability distribution of x?

The distribution of θ is rectangular, $dp = \pi^{-1}d\theta$. Also $x = \tan\theta$, so that $\theta = \arctan x$, $d\theta = dx/(1+x^2)$. Hence the distribution of x is given by

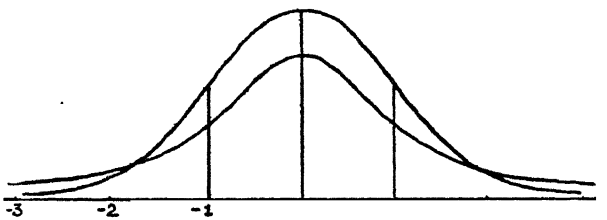
$$dp = \frac{1}{\pi}\frac{dx}{1+x^2}, \quad \text{range } -\infty \text{ to } \infty.$$

The probability function appearing here is Cauchy's probability function. It has the property (very awkward for any theory of estimation from sample based on moments) that its moments of even order $\mu_2, \mu_4, \ldots$ are all infinite. The reader should verify this by integration. It follows at once that linear compounding of independent variates obeying laws of Cauchy type cannot be carried out by the addition of seminvariants; in fact the seminvariant g.f.'s do not converge. This exception to the common rule gives us a salutary reminder that linear compounding of independent variates does not necessarily generate a distribution of normal type.

The Cauchy curve has been found to possess a specially remarkable property. If n independent variates obeying the same Cauchy law are added, and the mean is taken, this mean obeys exactly the same law. Not only so, but the distribution of any linear combination

$$z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$$

of variates x_j obeying the same Cauchy law, where the c_j are positive and sum to 1, is again exactly the same Cauchy distribution.



The figure shows the normal curve in standard measure and the flatter Cauchy curve drawn to the same scale.

39. The Pearson Curve of Type I. As a final example of a probability function arising from a particular problem, let us consider the following:

Suppose that x is distributed in the rectangular distribution over the range 0 to 1. Let $n+1$ points x_j be taken independently in this range. What is the probability that the $(k+1)^{th}$ point of these, as counted from the left of the range, is in the elementary interval $x-\frac{1}{2}dx$ to $x+\frac{1}{2}dx$?

The probability is compound; it is the probability that one, any one, of the $n+1$ points is in the interval, and that k of the remaining n are in the range 0 to $x-\frac{1}{2}dx$ while $n-k$ are in the range $x+\frac{1}{2}dx$ to 1. Hence the compound probability is

$$dp = \phi(x)dx = n_{(k)}(n+1)x^k(1-x)^{n-k}dx, \qquad (1)$$

for the first probability mentioned is $(n+1)dx$ and the second is $n_{(k)}x^k(1-x)^{n-k}$. The probability function $\phi(x)$ obtained here is of Pearson's Type I. It is in fact the integrand of the Beta function (Gillespie, *Integration*, p. 84), apart from the factor $n_{(k)}(n+1)$ which ensures that the

area under the curve is 1. Had the range been a to b, we should have obtained

$$\phi(x) = n_{(k)} \frac{n+1}{(b-a)^{n-}} (x-a)^k (b-x)^{n-k}. \qquad (2)$$

The probability integral of the simpler form (1) over the partial range (0, x) is called the *Incomplete Beta function.* In the same way the integral over (0, x) of the function

$$\phi(x) = \frac{1}{\Gamma(n)} x^{n-1} e^{-x}, \quad . \quad . \quad . \quad (3)$$

which is a case of Pearson's Type III, is called the *Incomplete Gamma function.*

Variety of Probability Curves. The preceding survey of types of probability function, though far from exhaustive, will have served to dispel the idea, once rather prevalent, that normality and symmetry were the rule and that skewness was an accident of sampling. The rôle of the normal distribution in statistics is not unlike that of the straight line in geometry; and we do not force curves into the mould of the straight line. Skew distributions are in fact the predominant type, for skewness arises from Lexian variability or non-homogeneity, from Poissonian statistical rarity, from limitation in the number of causes of variation, and from non-linear transformations of the scale.

CHAPTER IV

PRACTICAL CURVE-FITTING WITH STANDARD CURVES

40. Representation of Frequency Data by Normal Curve. The present chapter will be devoted to the numerical details of representing frequency distributions by normal curves, curves of Type A, Poissonian curves and curves of Type B.

In fitting the normal curve, that is, in finding the equation of the normal function of best approximation to the given frequency distribution, the idea is to represent the relative class frequencies by the corresponding segments of area under the normal curve between neighbouring ordinates corresponding to consecutive class boundaries. The mean m_1' or m of the frequency distribution is taken as the estimate of the mean μ or μ_1' of the normal function; the second moment m_2 or s^2, corrected for grouping if necessary by Sheppard's correction, is taken as the estimate of the corresponding μ_2 or σ^2. In order to use the standardized tables of the normal probability integral it is best, once m_1' and m_2 have been computed, to standardize the class boundaries, taking them as deviations from the mean, in units of s. The values of the probability integral corresponding to these class boundaries are then read from tables (Appendix 4); the first differences of these values are the estimates of the class probabilities; and finally we may multiply by n, the total number in sample, to make comparison with the absolute class frequencies.

Example. In the data of heights of Irishmen (**18**, Ex.) the mean is 67·34, and m_2 with Sheppard's correction is $4{\cdot}705 - 0{\cdot}083 = 4{\cdot}622$. Hence $s = 2{\cdot}15$, $1/s = 0{\cdot}465$. The standardized deviations of class boundaries are shown in the column $z = (x - m \pm \frac{1}{2})/s$ below. Since their common differ-

ence is $1/s$ or 0·465, they are readily found, when once any one of them has been computed, by repeated addition or subtraction of 0·465, and the results can be checked at the ends of the range. The next column shows the values of erf (z), the next the first differences of these, the next the same multiplied by 346, and the final column the original class frequencies themselves for comparison.

x	$z=(x-m\pm\frac{1}{2})/s$	$\frac{1}{2}$ erf z	$\frac{1}{2}\triangle$ erf z	$\frac{1}{2}n\triangle$ erf z	obs.
	$-\infty$	−0·5000			
59			0·0001	0	1
	−3·646	−0·4999			
60			0·0006	0	0
	−3·181	−0·4993			
61			0·0026	1	2
	−2·716	−0·4967			
62			0·0089	3	2
	−2·251	−0·4878			
63			0·0249	9	7
	−1·786	−0·4629			
64			0·0561	19	15
	−1·321	−0·4068			
65			0·1028	36	33
	−0·856	−0·3040			
66			0·1519	53	58
	−0·391	−0·1521			
67			0·1816	63	73
	0·074	0·0295			
68			0·1756	61	62
	0·539	0·2051			
69			0·1372	47	40
	1·004	0·3423			
70			0·0868	30	25
	1·469	0·4291			
71			0·0443	15	15
	1·934	0·4734			
72			0·0184	6	10
	2·399	0·4918			
73			0·0082	3	3
	∞	0·5000			
				346	346

41. Representation by Type A. The coefficients a_3, a_4, ... in the series **35** (8) of Type A can be expressed in terms of the moments about the mean. For by **35** (5) the m.g.f. (in unstandardized scale) is given by

$$1+\mu_2\alpha^2/2!+\mu_3\alpha^3/3!+\mu_4\alpha^4/4!+\ldots$$
$$=\exp(\tfrac{1}{2}\sigma^2\alpha^2)(1+a_3\sigma^3\alpha^3/3!+a_4\sigma^4\alpha^4/4!+\ldots). \qquad (1)$$

Multiply each of these expressions by $\exp(-\frac{1}{2}\sigma^2\alpha^2)$ and expand the product in the former case. Equating coefficients of $\alpha^r/r!$, we have the desired relations

$$\begin{aligned} a_3 &= \mu_3/\sigma^3, \\ a_4 &= (\mu_4-3\mu_2^2)/\sigma^4, \\ a_5 &= (\mu_5-10\mu_2\mu_3)/\sigma^5, \\ a_6 &= (\mu_6-15\mu_2\mu_4+30\mu_3^2)/\sigma^6, \qquad . \qquad . \qquad (2) \end{aligned}$$

and so on.

The routine for fitting Type A is a slight extension of that used in fitting the normal curve. Moments about the mean are computed and if necessary corrected by Sheppard's corrections. The coefficients a_3, a_4, ... are estimated from these moments by the formulæ just given, with m_r substituted for μ_r. The integral of the corresponding Type A series is then taken instead of the normal probability integral. This involves the necessity, if terms in a_3 and a_4 are included, of having supplementary tables of the integrals of the functions which appear in these terms, that is, tables of

$$-F_3(z)=\frac{1}{3!\sqrt{2\pi}}\left[\left(\frac{d}{dz}\right)^2 e^{-\frac{1}{2}z^2}\right]_0^z$$

and $$F_4(z)=\frac{1}{4!\sqrt{2\pi}}\left[\left(\frac{d}{dz}\right)^3 e^{-\frac{1}{2}z^2}\right]_0^z.$$

Such tables have been computed and are available. (British Association Tables, 1931 ; Bowley, *Elements of Statistics*, p. 303, $F_3(z)$ only.)

Example. (Bowley, *Elements of Statistics*, p. 309.) To

fit two terms of a series of Type A to data giving age distribution of St Louis school children in the sixth grade. (Age x means x to $x+1$.)

x	10	11	12	13	14	15	16	17	18	n
nf	26	201	673	1001	739	310	80	13	1	3044

By the usual routine we compute $m_1' = 13{\cdot}665$, $m_2 = 1{\cdot}498$, $m_3 = 0{\cdot}356$. Hence, using Sheppard's corrections, the corrected

$$s^2 = 1{\cdot}498 - 0{\cdot}083 = 1{\cdot}415, \quad s = 1{\cdot}190, \quad 1/s = 0{\cdot}840,$$

estimated $a_3 = m_3/s^3 = 0{\cdot}211$.

The rest of the working can be arranged in columns as below.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
x	$z=(x-m)/s$	$\frac{1}{2}\operatorname{erf} z$	$F_3(z)$	$a_3F_3(z)$	(3)+(5)	Δ	$n\Delta$	obs.	normal
10	$-\infty$	−0·5000	−0·0665	−0·0140	−0·5140				
						0·0079	24	26	38
11	−2·24	−0·4875	−0·0882	−0·0186	−0·5061				
						0·0678	206	201	208
12	−1·40	−0·4192	−0·0904	−0·0191	−0·4383				
						0·2202	670	673	630
13	−0·56	−0·2123	−0·0275	−0·0058	−0·2181				
						0·3268	995	1001	982
14	0·28	0·1103	−0·0076	−0·0016	0·1087				
						0·2440	743	739	786
15	1·12	0·3686	−0·0755	−0·0159	0·3527				
						0·1024	312	310	324
16	1·96	0·4750	−0·0942	−0·0199	0·4551				
						0·0264	80	80	68
17	2·80	0·4974	−0·0755	−0·0159	0·4815				
						0·0042	13	13	8
18	3·64	0·4999	−0·0672	−0·0142	0·4857				
						0·0003	1	1	0
19	∞	0·5000	−0·0565	−0·0140	0·4860		3044	3044	3044

The closeness to the observations is remarkable. Indeed the tests of "goodness of fit," to be developed in **54**, show that the discrepancies are so small as to be improbable, and the representation is unsatisfactory. We have here a case of "over-fitting."

For comparison we have included in a final column the results given by the normal curve of best agreement.

42. Representation by Poissonian Function or Type B. The coefficients b_2, b_3, ... in the series **35** (11)

of Type B can be expressed in terms of the factorial moments μ, $\mu_{(2)}$, $\mu_{(3)}$, ... For by **35** (10) the f.m.g.f. is given by

$$1+\mu\alpha+\mu_{(2)}\alpha^2/2!+\mu_{(3)}\alpha^3/3!+\ldots$$
$$=\exp(\mu\alpha)(1+b_2\alpha^2/2!+b_3\alpha^3/3!+\ldots) \quad . \quad . \quad (1)$$

Multiply each of these expressions by $\exp(-\mu\alpha)$ and expand the product in the former case. Equating coefficients of $\alpha^r/r!$ we have the desired relations

$$\begin{aligned} b_2 &= \mu_{(2)}-\mu^2, \\ b_3 &= \mu_{(3)}-3\mu_{(2)}\mu+2\mu^3, \\ b_4 &= \mu_{(4)}-4\mu_{(3)}\mu+6\mu_{(2)}\mu^2-3\mu^4, \end{aligned} \qquad (2)$$

and so on. Note that the numerical coefficients are the same as occur in **14** (5).

The procedure of fitting by Type B is therefore to compute factorial moments of the data by the summation method (Appendix **2**) and by substitution in the above formulæ to estimate the coefficients b_2, b_3, ... of the Type B series. For the rest of the work we require the values of $e^{-m}m^x/x!$ and its differences of as many orders as may be necessary.

The value of e^{-m} can be taken from a table of the exponential function. Then ne^{-m} is computed, after which each value of $ne^{-m}m^x/x!$ can be obtained from the preceding value, corresponding to $x-1$, by multiplying by m/x, most easily done by a calculating machine. The subsequent differencings and multiplication by coefficients b_2 and so on can best be followed from the illustrative example.

Example. E. Rutherford and H. Geiger, in 2608 experiments (*Phil. Mag.*, Ser. 6, **20**, 1910, p. 698) on the number x of α-particles radiated from a disc in 7·5 seconds, obtained the distribution :

x	0	1	2	3	4	5	6	7	8	9	10	11	12-14	n
nf	57	203	383	525	532	408	273	139	45	27	10	4	2	2608

The summation method for factorial moments gives $m = 3{\cdot}870$, $m_{(2)} = 14{\cdot}784$, whence the estimate of $b_2/2!$ is

$$\tfrac{1}{2}(14{\cdot}784 - 3{\cdot}87^2) = -0{\cdot}0965.$$

The working is set out in columns as below.

(1) x	(2) $n\psi$	(3) $n\triangle\psi$	(4) $n\triangle^2\psi$	(5) $\frac{1}{2}nb_2\triangle^2\psi$	(6) (2)+(5)	(7) obs.	(8) Poisson
0	54·40	54·40	54·40	−5·25	49	57	54
1	210·52	156·12	101·72	−9·82	201	203	211
2	407·37	196·85	40·73	−3·93	403	383	407
3	525·49	118·12	−78·73	7·60	533	525	526*
4	508·43	−17·06	−135·18	13·05	522	532	509*
5	393·52	−114·91	−97·85	9·44	403	408	394
6	253·81	−139·71	−24·80	2·39	256	273	254
7	140·34	−113·47	26·24	−2·53	138	139	140
8	67·89	−72·45	41·02	−3·96	64	45	68
9	29·18	−38·71	33·74	−3·26	26	27	29
10	11·29	−17·89	20·82	−2·01	9	10	11
11	3·96	−7·33	10·56	−1·02	3	4	4
12	1·28	−2·68	4·65	−0·45	1	2	1
13	0·39	−0·89	1·79	−0·17	0	0	0
					2608	2608	2608

N.B.—(i) In the differencings in columns (3) and (4) $\psi(-1)$, $\psi(-2)$... are tacitly taken as zero. (ii) The asterisked entries in column (8) have been raised from those in column (2) to make the totals of columns (7) and (8) both come to 2608.

It will appear when we come to consider goodness of fit (54) that the representation by the Poisson function alone, without the term in b_2, is satisfactory.

43. Limitations on the Use of Moments in Fitting Curves. The discussion of the Cauchy distribution in 38 has shown that moments are by no means always, or necessarily, the best parameters to use in representing an observed frequency distribution by a probability distribution of assigned functional form. It depends entirely on the nature of the probability function what parameters may be used with adequacy. For example, since the mean of any number of observations x, each of which obeys the same Cauchy distribution, has exactly the same Cauchy distribution as x, it follows that the mean of sample

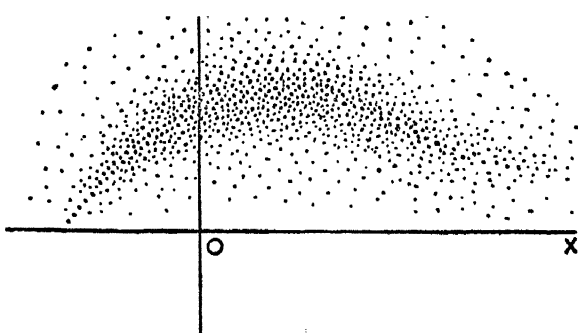
in this case is no more accurate, for the purpose of estimating the centroid of the curve, than any single observation; indeed it may be shown that the *median* is much superior for this purpose, while still better parameters can be found. Again, for the purpose of estimating the centre of an unknown rectangular probability distribution the mean of the n sample observations x_i is quite a good estimate; but surprisingly enough, as R. A. Fisher has shown, the *mean of the two extreme observations* alone is remarkably better. As a general precept it may be stated that for probability curves of shape and properties approximating to the normal curve the use of the mean and moments of the frequency distribution gives good estimates for those parameters in the probability distribution; but for other probability curves better parameters can be found.

CHAPTER V

PROBABILITY AND FREQUENCY IN TWO VARIATES

44. Bivariate Distributions: Correlation and Regression. Hitherto we have been concerned exclusively with probability and frequency distributions in one variate, that is, with univariate distributions. But most of the important and interesting applications of statistics involve bivariate, trivariate or multivariate distributions.

Let us consider how a typical bivariate frequency distribution may arise. Suppose that 1000 soldiers in a regiment are measured in height, x, and in weight, y. The measurements provide 1000 paired numbers (x_j, y_j), which may be plotted as points in a plane. The resulting assemblage of points may be called the "dot diagram."



Now there may be, and in fact in the case of height and weight there is, a tendency for the value of y_j to

conform in some way to that of the corresponding x_j; greater height as a rule is associated with greater weight. Any such tendency towards a functional relationship, obscured by random deviations, will manifest itself in the dot diagram by the greater density of the dots along a certain locus. This locus is not sharply outlined, but its estimation is important, for it is a smudged image of a curve which may be fine and clear-cut in the parent population of which the observations are a sample. This latent curve or functional relation $y = F(x)$ is called a *regression*, the *regression of y on x*. It will be a matter of judgement what functional basis is chosen for its mathematical representation. Usually the representation is a linear one based on a set of prescribed functions $p_1(x)$, $p_2(x)$, ..., the regression therefore appearing in the form

$$y = a_0 + a_1 p_1(x) + a_2 p_2(x) + \ldots, \quad . \quad . \quad (1)$$

to as many terms as are judged adequate. The statistical problem is then to determine the best estimates of a_0, a_1, a_2, ... from the n paired observations (x_j, y_j). The functions $p_i(x)$ are commonly polynomial or harmonic functions, but they may be of any preassigned functional type.

The diagram of dots suggests a second point of view. The proportion of dots in an elementary region $x - \frac{1}{2}\Delta x < x < x + \frac{1}{2}\Delta x, y - \frac{1}{2}\Delta y < y < y + \frac{1}{2}\Delta y$ gives an element of bivariate relative frequency which corresponds to a bivariate differential element of probability, let us say $dp = \phi(x, y)dxdy$, in the parent population.

We may imagine that on each class-rectangle of the network of rectangles delimited by class boundaries of x and y a right prism is erected, of volume proportional to the corresponding class frequency. The tops of these prisms make a surface of flat terraces which we may call the *prismogram*, the analogue in three dimensions of the histogram. This prismogram, then, is the rough sampling

approximation to an ideal probability surface $z = \phi(x, y)$, which is often called the *correlation surface.*

The functional dependence of y on x may be investigated either by the method of correlation, which consists in estimating the parameters of the bivariate probability function $\phi(x, y)$, or by the method of regression, which consists in estimating the coefficients a_j in the regression function (1). Naturally the methods overlap to a certain extent. In the case of several important correlation (bivariate probability) functions the corresponding regression curves are straight lines.

45. Binomial and Hypergeometric Correlation. The natural extension of the twofold division, success and failure of an event E, which gives rise to the binomial or hypergeometric distribution in one variate, is an arrangement giving a twofold division in each of two events E and F. Such an arrangement is expressed by the *fourfold table,* as follows:

Let the probabilities of the double events (E, F), $(E, \bar{F})$, $(\bar{E}, F)$ and $(\bar{E}, \bar{F})$ be p_{11}, p_{10}, p_{01}, p_{00}. These are set out as shown in the fourfold table, the columns referring

	E	$\bar{E}$	
F	p_{11}	p_{01}	p'
$\bar{F}$	p_{10}	p_{00}	q'
	p	q	1

to E and $\bar{E}$, the rows to F and $\bar{F}$. The sum $p_{11}+p_{10}$, representing the total probability of E whether F occurs or not, is entered marginally as p; and in the same way the other total probabilities q, p', q' are entered marginally as sums of a row or of a column.

Contingency Table. The fourfold table is a simple example of a *contingency table.* The more general bivariate contingency table has h rows and k columns, corresponding to the division of one system into h categories and the other into k. If the probabilities p_{ij} are all rational fractions, it is possible to represent the bivariate population by a physical model, such as one of marked or coloured balls in due proportions.

If E and F are independent events, then $p_{11} = pp'$, $p_{10} = pq'$, $p_{01} = qp'$ and $p_{00} = qq'$, so that $p_{11}p_{00} = p_{10}p_{01}$. The determinant $p_{11}p_{00} - p_{10}p_{01}$ of the fourfold table is thus zero.

Ex. 1. Prove that this determinant is equal to $p_{11} - pp'$ and to $p_{00} - qq'$.

Generating Functions. Just as in **7** we introduced a variable t to carry x as exponent in univariate generating functions, so it is natural to introduce u to carry y. The probability g.f. of a fourfold table will thus be

$$G(t, u) = p_{11}tu + p_{10}t + p_{01}u + p_{00} \quad . \quad . \quad . \quad (1)$$

$$= 1 + p(t-1) + p'(u-1) + p_{11}(t-1)(u-1). \quad . \quad (2)$$

Ex. 2. Show that in the case of independence this splits into the two factors $pt+q$, $p'u+q'$.

Now if we draw n times, with replacement each time, from the population characterized by the fourfold table, the g.f. will be

$$(p_{11}tu + p_{10}t + p_{01}u + p_{00})^n. \quad . \quad . \quad . \quad (3)$$

The coefficient of $t^x u^y$ in the expansion of this g.f. will be the probability $\phi(x, y)$ of having x cases E and y cases F in the n drawings. The function $\phi(x, y)$ is the correlation function of *binomial* type.

Ex. 3. If the variates x and y are independent, show that $\phi(x, y)$ is the simple product of the binomial probability functions

$$n_{(x)}p^x q^{n-x} \text{ and } n_{(y)}(p')^y (q')^{n-y}.$$

Again, if we have a fourfold population of N individuals with Np_{11}, Np_{10}, Np_{01} and Np_{00} individuals in the respective categories, and if we sample n times without replacement, the corresponding probability of x cases E and y cases F is the correlation function of *hypergeometric* type. If x and y are independent the function is the simple product of hypergeometric probability functions in x and y.

46. Bivariate Moments and Moment Generating Functions. The bivariate product moment of order r in x and s in y is defined by

$$\mu'_{rs} = \sum_{xy}\sum \phi(x, y)x^r y^s \quad \text{or} \quad \int\int \phi(x, y)x^r y^s dxdy, \quad . \quad (1)$$

or the corresponding mean values with $\sum_x \int dy$ or $\int dx \sum_y$, according to the discrete or continuous nature of the variables.

There are three moments of the second order. If we take them with respect to the means μ'_{10} and μ'_{01} of the variates they are μ_{20} the variance of x, μ_{02} the variance of y, and μ_{11} the *product moment* of x and y, often called the *covariance*.

Generating Functions. The bivariate generating function of probability is defined by

$$G(t, u) = \sum_{xy}\sum \phi(x, y)t^x u^y \quad \text{or} \quad \int\int \phi(x, y)t^x u^y dxdy, \qquad (2)$$

or the same with $\sum_x \int dy$ or $\int dx \sum_y$.

Moment generating functions are defined by putting $t = e^\alpha$, $u = e^\beta$, the general product moment μ'_{rs} being the coefficient of $\alpha^r \beta^s / r!\, s!$ in the resulting m.g.f.

Factorial moments can be defined by putting factorials $x^{(r)}$ and $y^{(s)}$, as defined in **29** (2), instead of powers x^r and y^s; and a bivariate f.m.g.f. may be constructed by putting $t = 1+\alpha$, $u = 1+\beta$.

Example. Prove that the f.m.g.f. of the fourfold table is

$$(1+p\alpha)(1+p'\beta)+(p_{11}p_{00}-p_{10}p_{01})\alpha\beta.$$

47. Normal Correlation as the Limit of Binomial Correlation. The g.f. of a sample of n drawings, with replacement each time, from the fourfold population is

$$(p_{11}tu+p_{10}t+p_{01}u+p_{00})^n$$
$$[1+p(t-1)+p'(u-1)+p_{11}(t-1)(u-1)]^n, \qquad (1)$$

by **45** (2).

Just as at the corresponding stage in **31**, let us consider the deviation of *relative* frequency of number of successes from means, rather than absolute frequency. We do this by putting $t = e^{\alpha}$, $u = e\beta$ in (1) and then writing α/n for α, β/n for β. We have then the bivariate m.g.f.

$$[1+p\alpha/n+p\alpha^2/2n^2+p'\beta/n+p'\beta^2/2n^2+p_{11}\alpha\beta/n^2+O(n^{-3})]^n$$
$$=[(1+p\alpha/n+\tfrac{1}{2}p^2\alpha^2/n^2)(1+p'\beta/n+\tfrac{1}{2}p'^2\beta^2/n^2)(1+\tfrac{1}{2}(p-p^2)\alpha^2/n^2$$
$$+\tfrac{1}{2}(p'-p'^2)\beta^2/n^2+(p_{11}-pp')\alpha\beta/n^2+O(n^{-3})]^n, \quad . \quad . \quad . \quad (2)$$

which tends asymptotically as n increases (the assumption throughout being that none of the probabilities in the fourfold table is $O(n^{-1})$) to

$$\exp(p\alpha+p'\beta)\exp\tfrac{1}{2}(\sigma_1^2\alpha^2+2\rho\sigma_1\sigma_2\alpha\beta+\sigma_2^2\beta^2) \qquad (3)$$

where $\sigma_1^2 = pq/n$, $\sigma_2^2 = p'q'/n$, $\rho\sigma_1\sigma_2 = (p_{11}-pp')/n$.

Next, just as in the case of one variate treated in **31**, and for analogous reasons, the question is to find a continuous function $\phi(x, y)$ satisfying

$$\int_{-\infty}^{\infty}\int \phi(x,y)e^{\alpha x+\beta y}dxdy = \exp\tfrac{1}{2}(\sigma_1^2\alpha^2+2\rho\sigma_1\sigma_2\alpha\beta+\sigma_2^2\beta^2). \quad (4)$$

The answer provided by pure mathematics is that the

only function $\phi(x, y)$ for which this is the case over a finite domain in α and β together is

$$\phi(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\tfrac{1}{2}Q(x, y)\right],$$

$$\text{where } Q(x, y) = (x^2/\sigma_1^2 - 2\rho xy/\sigma_1\sigma_2 + y^2/\sigma_2^2)/(1-\rho^2). \quad . \quad (5)$$

This function, the analogy of which with the normal probability function in one variable is evident, is the *bivariate normal probability function* or *normal correlation function*. The parameter ρ is called the *coefficient of correlation*. The reader will verify at once that when $\rho = 0$ the correlation function breaks up, as one might have expected, into the product of two ordinary normal functions, in x and y respectively.

48. Properties of the Normal Correlation Function. Let us suppose that units of scale in x and y are standardized by putting $\sigma_1 = 1$, $\sigma_2 = 1$. The m.g.f. of the normal correlation function about the means then becomes

$$\exp \tfrac{1}{2}(\alpha^2 + 2\rho\alpha\beta + \beta^2), \quad . \quad . \quad . \quad (1)$$

and the coefficient of $\alpha\beta/1!\ 1!$ in the expansion of this shows that ρ is the mean value of the product xy. This suggests that in computing the parameters of bivariate *frequency* distributions we should add to the usual four parameters of first and second order, namely the means m'_{10}, m'_{01} and variances s_1^2 and s_2^2 of x and y, a fifth parameter, the mean value of the *product* of corresponding deviations x and y from the sample means.

The standardized value of this mean product, namely,

$$r = \frac{1}{N}\Sigma(x - m'_{10})(y - m'_{01})/s_1 s_2, \quad (2)$$

corresponds in the sample to ρ in the population or probability function. We shall call r the *Pearsonian* coefficient, or *product-moment* coefficient, of x and y in the frequency distribution.

Limits of r and ρ. The extreme values that r and ρ can take are 1 and -1. They cannot lie outside those limits. For, taking x and y as unstandardized deviations from their means, let us consider the mean value of $(hx+ky)^2$, or $h^2x^2+2hkxy+k^2y^2$, both in population and in sample, where h and k are arbitrary. In population the mean value is $h^2\sigma_1^2+2hk\rho\sigma_1\sigma_2+k^2\sigma_2^2$, in sample it is $h^2s_1^2+2hkrs_1s_2+k^2s_2^2$. . . . . (3)

Now these quadratic expressions in h and k, being the mean values of squared functions, are of necessity not negative. But the necessary condition for this is that the discriminants

$$(\rho\sigma_1\sigma_2)^2-\sigma_1^2\sigma_2^2 \quad \text{and} \quad (rs_1s_2)^2-s_1^2s_2^2 \; . \qquad (4)$$

should not be positive. Hence $\rho^2\leqslant 1$, $r^2\leqslant 1$, so that both ρ and r must lie in the range -1 to 1.

The result, it may be noted, depends on a property of quadratic expressions, and therefore holds not merely for normal but for *any* distribution of x and y.

Example. Prove that if x and y are uncorrelated and of unit variance, $x\cos\theta+y\sin\theta$ and $x\sin\theta-y\cos\theta$ are also uncorrelated and of unit variance.

In the case of independent variates, under any laws of distribution, the product moment μ_{11} about the means is zero. For if the separate m.g.f.'s of x and y about their means are

$$1+\mu_{20}\alpha^2/2!+\ldots \quad \text{and} \quad 1+\mu_{02}\beta^2/2!+\ldots, \qquad (5)$$

then by compound probability the m.g.f. of the two together is

$$(1+\mu_{20}\alpha^2/2!+\ldots)(1+\mu_{02}\beta^2/2!+\ldots), \qquad (6)$$

and since this has no term in $\alpha\beta$ we have $\mu_{11}=0$.

It is most important to notice that the converse theorem is not true. *The vanishing of μ_{11} does not imply independence.* Consider for example the case when x is distributed in any symmetrical distribution about the

mean $x = 0$, with variance σ^2. Then $z = x^2 - \sigma^2$ is also distributed about a zero mean. The variates x and z have complete functional dependence. Their product xz however is $x^3 - \sigma^2 x$, and the mean value of this is clearly zero. This is an extreme case, but it gives a sharp warning against inferring the existence of independence from a zero value of ρ, and still more from a zero value of r, which is merely an estimate of ρ.

The normal correlation surface, when $\rho = 0$ and variances are standardized to unity, is a symmetrical bell-shaped surface which may be generated by the rotation of its central vertical section, a normal curve, about the vertical axis. When $\rho \neq 0$ the surface acquires a hog-back ridge which lies in the first and third quadrants of (x, y) if ρ is positive, in the second and fourth quadrants if ρ is negative.

The loci of equal *probability density* (**7**) are found by equating $\phi(x, y)$ to a constant, yielding curves of the form

$$x^2/\sigma_1^2 - 2\rho xy/\sigma_1\sigma_2 + y^2/\sigma_2^2 = c^2. \quad . \quad . \quad (7)$$

These are homothetic ellipses. Among them the ellipse which includes a region in x and y of total probability $\frac{1}{2}$ is sometimes called the "probable ellipse," a name, like "probable error" in **15**, apt to mislead. This region is the bivariate analogue of the interquartile range (**15**).

49. Regression Lines in Bivariate Normal Correlation. If we cut the normal correlation surface by a series of planes all perpendicular to the axis of x, the sections are all normal curves. For each such section corresponds to a constant value of x_k of x, and so the z-ordinate of such a section is, in standard scale,

$$z = \phi(x_k, y) = c \exp\left[-\tfrac{1}{2}(x_k^2 - 2\rho x_k y + y^2)/(1-\rho^2)\right], \quad . \quad (1)$$

$$= c_1 \exp\left[-\tfrac{1}{2}(y - \rho x_k)^2/(1-\rho^2)\right], \quad . \quad . \quad (2)$$

where c and c_1 are constants; and this is the ordinate

of a normal curve with mean at $y = \rho x_k$ and of variance $1-\rho^2$, or in unstandardized scale $\sigma_2^2(1-\rho^2)$; this variance is the same for all such sections. The locus of the means of such sections is therefore the straight line $y = \rho x$, or in unstandardized units $y/\sigma_2 = \rho x/\sigma_1$. This straight line is the regression line of y on x. There is correspondingly a regression line $x = \rho y$, or $x/\sigma_1 = \rho y/\sigma_2$, of x on y.

The regression lines do not coincide unless $\rho = \pm 1$, in which case (with standard units) they are the bisectors of the angles between the x and y axes. If $\rho = 0$ the regression lines are the axes themselves; but the concept of regression is of little importance in this case.

Note. The name "regression" was introduced by Sir Francis Galton (*J. Anthrop. Inst.*, 15 (1886), p. 246). In bivariate data concerning heights of fathers, x_j, and heights of eldest sons, y_j, he found that the regression lines, as estimated from the sample, were approximately $y = \frac{1}{2}x, x = \frac{1}{2}y$. This implies, for example, that if there is a group of fathers whose heights *all* deviate from mean height by d inches, then the *average deviation* of the height of their sons from mean height is only $\frac{1}{2}d$. There is thus a tendency, in the next generation, to return or *regress* towards the mean. If this feature of regression were not present, a character such as height might acquire greater and greater dispersion in succeeding generations.

50. Correlation Table: Computation of Product-Moment. A contingency table of h rows and k columns in which both variables x and y are metrical is called a *correlation table.* If x and y are continuous variates it will be convenient to take a class-unit of suitable size for each and thus to have class-frequencies corresponding to class-rectangles. For practical purposes it is advisable to choose these units so that each variate has ten or a dozen classes, not more.

The following example illustrates the usual appearance of a correlation table. (The distribution is of Binet Intelligence Quotient, x, and Verbal Score, y, of 500 Scottish

schoolgirls born in 1921, tested in the first week of June 1932. *The Intelligence of Scottish Children*, Univ. of London Press, 1933, p. 96.) The score named 60 means 60 and over, that is, the class 60 to 69, so that the class marked 60 in the report should be centred at 64·5; and so for other classes. The sums of rows and columns are entered in the margins; they give the frequency distribution of x when variation in y is neglected, and of y when variation in x is neglected.

		x (Binet I.Q.)										
		60	70	80	90	100	110	120	130	140	150	f_y
	70								2			2
	60					3	2	6	3	4	1	19
y	50				10	15	26	19	14	2		86
(Verbal	40		2	7	32	43	23	7	2	0	1	117
Score)	30		2	28	50	31	15	2	1			129
	20		10	32	38	6	1					87
	10		11	28	4							43
	0	3	7	7								17
	f_x	3	32	102	134	98	67	34	22	6	2	500

From the marginal distributions we can proceed to compute the means and mean-square-deviation from means of x and y. This will always be the first step in computing r. The product-moment can be computed about provisional means, and then transferred by a correction to the true means, thus:

$$\text{Since } \frac{1}{N}\Sigma x = m'_{10}, \frac{1}{N}\Sigma y = m'_{01},$$

the product-moment about these means is

$$\frac{1}{N}\sum_x\sum_y(x-m'_{10})(y-m'_{01}) = \frac{1}{N}(\Sigma\Sigma xy - m'_{10}\Sigma y - m'_{01}\Sigma x) + m'_{10}m'_{01}$$

$$\frac{1}{N}\Sigma\Sigma xy - m'_{10}m'_{01}. \quad (1)$$

Hence, just as mean-square-deviations can be com-

puted about a provisional mean and transferred (**14**) to the true mean by subtracting a *square*, so mean-product-deviations can be so transferred by subtracting the corresponding *product* of deviations of provisional means. It is to be observed that this product may be negative, in which case the correction involves an addition.

Several different methods of computation are in use for finding r. We shall exemplify two, of which the rest are mostly variants.

(i) The first method consists in computing $\Sigma\Sigma xy$ piecemeal according to the contributions made to this sum by the frequencies in the rows, or alternatively in the columns. For example, in the k^{th} row, for $y = y_k$ constant, we compute $\sum_j f_j x_j$, that is, multiply each class-frequency f_j by the value of x, x_j, and add along the row. For the different rows we may enter these values $\sum_j f_j x_j$ in a suitable column to the right. The sums of such values for all rows gives Σx, and so may be used to check the mean m'_{10}; while if we multiply each entry in that added column by its appropriate y_k and sum down the column we have $\Sigma\Sigma xy$.

The same procedure may be carried out by columns instead of rows. We then have a check on both means and on $\Sigma\Sigma xy$. The whole scheme can be neatly arranged in rows and columns annexed to the table as below. The special value of the arrangement is perceived when it is found necessary to compute correlation ratios (**52**) as well as correlation coefficients. It simplifies the arithmetic, too, to choose units such that the class-breadths of x and y are both unity.

Ex. 1. By way of explanation of the entries, note that the second entry, 63, in the Σx column comes from $3\times1+2\times2+6\times3+3\times4+4\times5+1\times6$, while the second entry, -51, in the Σy row comes from

$$2\times1+2\times0+10\times(-1)+11\times(-2)+7\times(-3).$$

y	−3	−2	−1	0	1	2	3	4	5	6	f_y	y	yf	y^2f	Σx	$y\Sigma x$
4								2			2	4	8	32	8	32
3					3	2	6	3	4	1	19	3	57	171	63	189
2				10	15	26	19	14	2		86	2	172	344	190	380
1		2	7	32	43	23	7	2	0	1	117	1	117	117	113	113
0		2	28	50	31	15	2	1			129	0	0	0	39	0
−1		10	32	38	6	1					87	−1	−87	87	−44	44
−2		11	28	4							43	−2	−86	172	−50	100
−3	3	7	7								17	−3	−51	153	−30	90
f_x	3	32	102	134	98	67	34	22	6	2	500		130	1076	289	948
x	−3	−2	−1	0	1	2	3	4	5	6						
xf	−9	−64	−102	0	98	134	102	88	30	12	289					
x^2f	27	128	102	0	98	268	306	352	150	72	1503					
Σy	−9	−51	−102	6	76	80	63	47	16	4	130					
$x\Sigma y$	27	102	102	0	76	160	189	188	80	24	948					

$m'_{10} = 289/500 = 0{\cdot}578.$

$m'_{01} = 130/500 = 0{\cdot}260.$

$s_1^2 = 1503/500 - (0{\cdot}578)^2 = 2{\cdot}672.$

$s_2^2 = 1076/500 - (0{\cdot}260)^2 = 2{\cdot}084.$

$rs_1s_2 = 948/500 - (0{\cdot}578)(0{\cdot}260) = 1{\cdot}746.$

Hence

$$s_1 = 1{\cdot}635, \quad s_2 = 1{\cdot}444, \quad r = 1{\cdot}746/1{\cdot}635 \times 1{\cdot}444 = +0{\cdot}74.$$

Error of Sampling.* The coefficient r, computed in this way, has a probability distribution (Chapter VII) depending on the probability distribution of x and y and on the number n. If the distribution of x and y is normal the sampling distribution of r tends with increasing n to become a normal distribution with variance $(1-\rho^2)^2/(n-1)$. Consequently the standard deviation of the sampling distribution of r (" standard error " of r) is approximately $(1-r^2)/\sqrt{n}$, but this is only so when n is large, let us say $n>100$, and when $|\rho|$ is not too high, let us say not greater than 0·5. In fact it is better in most cases, and certainly when n is small, to estimate by tables of R. A. Fisher's distribution of r within what range r may be taken as an estimate of ρ.

(ii) The second method of computing r depends on the simple observation that while by summing the frequencies in columns we obtain the distribution in x alone, and by rows that in y alone, if we sum along diagonals inclined at 45° to the horizontal we obtain a

* This paragraph may be postponed until Chapter VII has been studied.

distribution of $x-y$; for all class-rectangles in any such diagonal correspond to the same value of $x-y$. Thus from diagonal frequencies we may compute the mean of $x-y$, namely $m'_{10}-m'_{01}$, thereby checking the individual means as computed from row and column marginal frequencies; and we may also compute the mean-square-deviation of $x-y$ from its mean. Now the value of this is

$$(x-y)^2-(m'_{10}-m'_{01})^2 = s_1^2-2rs_1s_2+s_2^2.$$

But s_1^2 and s_2^2 are already known from the row and column marginal distributions; hence r is easily found.

Ex. 2. Taking the same example as before and summing along the diagonals, we find the frequency distribution of $x-y$ to be

$x-y$	-3	-2	-1	0	1	2	3	4	5	N
Nf	2	22	87	173	149	57	8	1	1	500

The mean is found to be 0·318, checking the values $m'_{10} = 0{\cdot}578$, $m'_{01} = 0{\cdot}260$. The mean-square-deviation from the mean is

$$s_1^2-2rs_1s_2+s_2^2 = 1{\cdot}265,$$

whence

$$\begin{aligned} r &= \tfrac{1}{2}(2{\cdot}672+2{\cdot}084-1{\cdot}265)/1{\cdot}635\times 1{\cdot}444 \\ &= \tfrac{1}{2}\times 3{\cdot}491/2{\cdot}361 = 0{\cdot}74, \end{aligned}$$

as before.

Notice that we have here no check on r. That could be provided by summing along the other set of diagonals at right angles to those which have been taken. They correspond to constant values of $x+y$, and so their mean-square- deviation from the mean is $s_1^2+2rs_1s_2+s_2^2$.

Ex. 3. The distribution of $x+y$, obtained by summing along the other set of diagonals in the correlation table, is:

$x+y$	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	N
Nf	3	7	18	38	38	68	63	64	68	40	37	23	20	6	6	1	500

Compute r from this distribution. Notice how much more widely spread it is than that of $x-y$ in Ex. 2.

Sheppard's Corrections. Sheppard's correction for variance in grouped data is applicable to the mean-square-deviations of x and y, but not to the mean-product-deviation. On the whole, however, it is better to work without the corrections, because the tables of Fisher's sampling distribution of r do not take account of grouping.

51. Correlation of Variates with Poissonian Distribution. It is not necessarily true that sampling from a fourfold population always produces as a limiting case a bivariate normal correlation function. Suppose, for example, that p and p' are of order $1/n$. Then p_{11} may be of order $1/n^2$, but may also be of order $1/n$.

The f.m.g.f. of a sample of n individuals with replacement is seen from **47** (1) to be

$$(1+p\alpha+p'\beta+p_{11}\alpha\beta)^n. \qquad (1)$$

When $p = \mu/n$ and $p' = \mu'/n$, and p_{11} is $O(1/n^2)$, this g.f. tends to $\exp(\mu\alpha+\mu'\beta)$, which shows that with increasing n the probability function reduces to the product of independent Poissonian functions, and is in fact

$$\psi(x, y) \quad : e^{-\mu}\frac{\mu^x}{x!}\, e^{-\mu'}\frac{(\mu')^y}{y!} \qquad (2)$$

On the other hand, when p_{11} is $O(1/n)$, we have

$$(1+p\alpha+p'\beta+p_{11}\alpha\beta)^n$$
$$= [(1+p\alpha)(1+p'\beta)(1+\overline{p_{11}-pp'}\alpha\beta+O(n^{-2}))]^n,$$

which tends to

$$\exp(\mu\alpha+\mu'\beta+\bar{\mu}\alpha\beta), \quad . \quad . \quad . \quad (3)$$

where

$$\bar{\mu} = n(p_{11}-pp') = n(p_{11}p_{00}-p_{10}p_{01}). \quad . \quad . \quad (4)$$

Evidently $\bar{\mu}$ is the ordinary product-moment about the means.

Now putting $\alpha = t-1$, $\beta = u-1$ in (3), we derive the

correlation function $\psi(x, y)$ as the coefficient of $t^x u^y$ in the probability g.f.

$$e^{-(\mu+\mu'-\bar{\mu})}e^{(\mu-\bar{\mu})t+(\mu'-\bar{\mu})u+\bar{\mu}tu} \tag{5}$$

It is found without difficulty to be

$$\psi(x, y) = e^{-(\mu+\mu'-\bar{\mu})} \frac{(\mu-\bar{\mu})^x(\mu'-\bar{\mu})^y}{x!\ y!}$$

$$\times \left|1+ \frac{\bar{\mu}xy}{(\mu-\bar{\mu})(\mu'-\bar{\mu})} + \frac{\bar{\mu}^2x(x-1)y(y-1)}{(\mu-\bar{\mu})^2(\mu'-\bar{\mu})^2 2!} + \right. \tag{6}$$

where the polynomial in the bracket terminates after $x+1$ or $y+1$ terms, whichever is the lesser. This function is the bivariate Poissonian function. It may be proved that the loci of means of sections corresponding to constant x or y are straight lines, so that here again we have linear regression. The same property may be proved to hold for binomial and hypergeometric correlation functions.

Both the normal function and the Poissonian correlation function can be derived, like the corresponding functions for one variate, on more general grounds than sampling from a fourfold table, by a compounding of elementary increments achieved by addition of bivariate seminvariant g.f.'s ; but this derivation lies beyond our present scope.

52. Non-Linear Correlation and Regression. A linear regression between correlated variates is rather exceptional. The loci of means of arrays usually deviate from straightness by more than can be ascribed to random sampling, suggesting that the underlying law of probability cannot be either normal or Poissonian. Non-linear regression curves are perhaps best estimated by fitting to the data suitable regression functions by the method of Least Squares, described in Chapter VI. In the non-linear case, too, the coefficient r or ρ has marked disadvantages (it was seen for example in **48** that ρ could be zero even when regression was perfect) and the correlation ratio η, devised by K. Pearson, is much to be preferred.

It was proved in 49 that in normal regression all y-sections or arrays corresponding to constant x had the same variance, let us say

$$\sigma_{2,1}^2 = \sigma_2^2(1-\rho^2), \quad . \quad . \quad . \quad (1)$$

so that

$$1-\rho^2 = \sigma_{2,1}^2/\sigma_2^2. \quad . \quad . \quad . \quad (2)$$

Now $1-\rho^2$ may be regarded as measuring something complementary or antithetic to correlation. The word *alienation* is sometimes used to describe this quality, but alienation suggests repulsion and is too strong a term. *Residual dispersion* expresses the meaning better. In non-linear regression the variance of the y-sections, namely

$$\sigma_{2,x}^2 = \int (y-\bar{y}_x)^2\phi(x, y)dy \Big/ \int \phi(x, y)dy, \quad . \quad . \quad (3)$$

where

$$\bar{y}_x = \int y\phi(x, y)dy \Big/ \int \phi(x, y)dy, \quad . \quad . \quad . \quad (4)$$

the mean of the y-section corresponding to constant x, is not usually constant. We may, however, take the mean of these variances of y-sections over all sections, that is, over all values of x, obtaining

$$\sigma_{2,1}^2 = \iint (y-\bar{y}_x)^2\phi(x, y)dxdy, \quad . \quad (5)$$

which may be regarded as the mean-square-deviation of y from its regression value $\bar{y}_x$, taken over the whole distribution. Standardizing this by dividing by the total variance of y, namely σ_2^2, and writing

$$1-\eta_{yx}^2 = \sigma_{2,1}^2/\sigma_2^2, \quad . \quad . \quad . \quad (6)$$

we define a coefficient η_{yx} analogous to ρ in (2). This coefficient η_{yx} is the *correlation-ratio of y on x*. The closer it approaches 1, the smaller is the residual dispersion and the closer the values y lie to their regressional means.

In the same way, by interchanging the rôles of x and y in the above derivation, we define η_{xy}, the correlation-ratio of y on x. As to the signs of η_{yx} and η_{xy}, there are cases where these can be attributed by graphical or other considerations, but there are also cases, for example when the curve of regression is a periodic curve with several oscillations, when sign has no meaning.

The estimates of η_{yx} and η_{xy}, as derived from an actual frequency distribution presented as a correlation table, will be denoted by e_{yx} and e_{xy}. We define them analogously; thus

$$1-e_{yx}^2 = s_{2,1}^2/s_2^2, \qquad . \qquad . \qquad . \quad (7)$$

where $s_{2,1}^2$ is the mean, over all y-arrays (columns of the correlation table), of the mean-square-deviation of y from the mean $\bar{y}_x$ of the column. In computing this mean of mean-square-deviations the column frequencies, marginally entered, serve as class frequencies. The effective arithmetical arrangement of the computation will be given later.

That the correlation-ratio is actually a ratio, namely the ratio of the standard deviation of the means of arrays to the total standard deviation of the variate, will now be proved by considering e_{yx}.

Lemma. If k sets of $n_1, n_2, \ldots, n_k$ observations, with respective means M_j and mean-square-deviations s_j^2, $j = 1, 2, \ldots, k$, are pooled in an aggregate of $n = n_1+n_2+\ldots+n_k$ observations with mean M and mean-square-deviation s^2, then

$$ns^2 = \sum_j n_j(s_j^2+c_j^2), \qquad . \qquad . \quad (8)$$

where $c_j = M-M_j$.

This follows at once from the fact that the mean-square-deviation of the j^{th} set about M is $s_j^2+c_j^2$.

Applying this lemma to the column-arrays of a correlation table, we have

$$ns_2^2 = \Sigma n_j(s_j^2+M_j^2), \qquad (9)$$

where M_j and s_j^2 are the mean and mean-square-deviation of the j^{th} column. (The origin is the mean of both x and y.) This is the same as

$$s_2^2 = s_{2,1}^2 + s_{M_j}^2, \quad . \quad . \quad . \quad (10)$$

the second term denoting the mean-square-deviation of column means, when these are associated with column frequencies.

The above result holds for the sample. A similar result can be proved for in the population, integrals replacing sums, and variance replacing mean-square-deviation from means. The result may be put in words thus : *total variance of* y *is equal to mean of variances* $\sigma_{2,x}^2$ *of* y*-arrays plus variance* $\sigma_{M_x}^2$ *of means of arrays.* It is another example of analysis of variance (**26, 75**).

Hence, by (6) and (7),

$$\eta_{yx}^2 = \sigma_{M_x}^2/\sigma_2^2 \quad \text{and} \quad e_{yx}^2 = s_{M_j}^2/s_2^2, \quad . \quad . \quad (11)$$

so that η_{yx}^2 and e_{yx}^2 are displayed as ratios of variances or of mean-square-deviations.

53. Computation of Correlation-Ratios. The result **52** (11) permits us to compute e_{yx} and e_{xy} by a simple extension of the first method of **50** for computing r, for the means of rows and columns are given by the entries in the column headed Σx and the row headed Σy, divided respectively by the frequencies f_y and f_x. Also, the means of these entries are m'_{10} and m'_{01}. Hence, computing mean-square-deviations from means in the usual way, we have

$$\begin{aligned} , &= \left[\frac{1}{N}\Sigma f_x(\Sigma y/f_x)^2 - m'^2_{01}\right]/s_2^2 \\ &= \left[\frac{1}{N}\Sigma(\Sigma y)^2/f_x - m'^2_{01}\right]/s_2^2, \end{aligned} \quad (1)$$

and similarly for e_{xy}^2. We thus annex two rows $(\Sigma y)^2$, $(\Sigma y)^2/f_x$, and two columns $(\Sigma x)^2$, $(\Sigma x)^2 f_y$, to the computation scheme for r.

Example. The additional rows and columns for the example of **50** (Binet I.Q. and Verbal Test Score) are as follows :

$(\Sigma y)^2$	81	2601	10404	36	5776	6400	3969	2209	256	16		$(\Sigma x)^2$	$(\Sigma x)^2/f_y$
$(\Sigma y)^2/f_x$	27	81	102	0	59	96	117	100	43	8	633	64	32
												3969	209
												36100	420
												12769	109
												1521	12
												1936	22
												2500	58
												900	53
													915

$$e^2_{yx} = [633/500 - (0{\cdot}260)^2]/(1{\cdot}444)^2$$
$$= 1{\cdot}198/2{\cdot}085 = 0{\cdot}575$$
$$e^2_{xy} = [915/500 - (0{\cdot}578)^2]/(1{\cdot}635)^2$$
$$= 1{\cdot}496/2{\cdot}673 = 0{\cdot}560.$$

Hence e_{yx} and e_{xy} are equal to 0·76 and 0·75, whereas the value of r was found to be 0·74.

54. Correlation of Non-Metrical Characters. When the characters in a double classification are purely qualitative, capable of being graded by a recognizable difference in category, but not susceptible of measurement by metrical scale, we must fall back on the contingency table of $h \times k$ rectangular cells, with corresponding cell-frequencies. Since variances and product-moments are now out of the question, the presence or absence of correlation must be inferred from the cell-frequencies themselves, according to the manner in which they deviate from presumptive cell-frequencies in the corresponding case of independence.

Consider, for example, the following contingency table due to Galton (*Proc. Roy. Soc.*, 40 (1886), p. 42), illustrating the incidence of eye-colour in a group of fathers and eldest sons.

(i)

	E_1	E_2	E_3	E_4	p'
F_1	0·194	0·083	0·025	0·056	0·358
F_2	0·070	0·124	0·034	0·036	0·264
F_3	0·041	0·041	0·055	0·043	0·180
F_4	0·030	0·036	0·023	0·109	0·198
p	0·335	0·284	0·137	0·244	1·000

		E_1	E_2	E_3	E_4	p'
	F_1	0·120	0·102	0·049	0·087	0·358
(ii)	F_2	0·089	0·075	0·036	0·064	0·264
	F_3	0·060	0·051	0·025	0·044	0·180
	F_4	0·066	0·056	0·027	0·049	0·198
		0·335	0·284	0·137	0·244	1·000

The colour categories are 1, blue; 2, blue-green or grey; 3, dark grey or hazel; 4, brown. $n = 1000$.

Summing down columns we obtain frequency estimates of the probabilities p of respective eye-colours for fathers irrespective of sons, and summing along rows, frequency estimates of probabilities p' for sons alone. These marginal frequencies or relative frequencies may be recombined again to form a multiplication table, which is to serve for comparison with the original table. The marginal frequencies in the second table are the same as in the first, but the cell-frequencies, derived as they are by applying the law of compound probability, represent what would have been the state of affairs with the same marginal frequencies had there been independence. Of course it must be observed that if we use, as here, not the *à priori* marginal probabilities but only the sample estimates given by the marginal frequencies, this procedure is bound to affect the sampling probability of the coefficient or criterion of comparison, χ^2.

The coefficient χ^2 is a quadratic function of the deviations of cell-frequencies in the actual from those in the presumptive independent case; it is a kind of composite weighted variance, with application not merely to contingency tables but also to any comparison of actual frequency classifications, single or multiple, with presumptive ones. It was first employed by Lexis, but the nature of its probability distribution was first obtained by K. Pearson in 1900.

Distribution of χ^2. The derivation of the χ^2-distribution involves the general multivariate normal correlation function, which is outside the scope of this short book; but the outlines may be sketched. If the *à priori* probabilities in the k classes are $p_1, p_2, \ldots, p_k$, then the frequencies in the classes are characterized by the multivariate multinomial g.f.

$$(p_1t_1+p_2t_2+\ldots+p_kt_k)^n. \quad . \quad . \quad . \quad (1)$$

If the number of individuals found in a class is n_j, the expected number being np_j, we may denote the class deviation from mean value or expectation n_j-np_j by ϵ_j. Then, since $\Sigma n_j = n$ and also $\Sigma np_j = n$, we must have

$$\Sigma\epsilon_j = 0, \quad . \quad . \quad . \quad . \quad (2)$$

a relation in virtue of which only $k-1$ of the deviations ϵ_j, let us say the first $k-1$, are independent. We therefore put $t_k = 1$ in the g.f. and consider what happens as n increases. Putting $t_j = e^{a_j}$, we find that, provided no p_j is $O(n^{-1})$, the multivariate m.g.f. of the class deviations tends to

$$\exp\left[\tfrac{1}{2}n \sum_{i,j=1}^{k-1}(p_jq_ja_j^2-2p_ip_ja_ia_j)\right] \quad . \quad . \quad (3)$$

This is an m.g.f. of normal correlated probability in the $k-1$ deviations, which on reversion gives the probability differential of the ϵ_j as

$$c \exp\left[-\tfrac{1}{2}n^{-1}\sum_{j=1}^{k}\epsilon_j^2/p_j\right]d\epsilon_1 d\epsilon_2 \ldots d\epsilon_{k-1}. \quad . \quad . \quad (4)$$

Thus the probability, or probability density, of a set ϵ_j of deviations is a function of the quadratic expression

$$\chi^2 = \Sigma\epsilon_j^2/np_j, \quad . \quad . \quad . \quad (5)$$

which is Pearson's χ^2. Having decided to use the composite χ^2 rather than the individual deviations ϵ_j as a criterion of the nearness to expectation, we transform the

differential (4) into a differential in χ^2 itself, when it assumes the shape

$$dp = c\chi^{k-3}e^{-\frac{1}{2}\chi^2}d\chi^2, \quad . \quad . \quad . \quad (6)$$

the probability function being of Pearson's Type III or Gamma type.

The probability of obtaining a value of χ^2 not exceeding a given χ_0^2 is therefore

$$P_{\chi_0^2} = \int_0^{\chi_0^2} \chi^{k-3}e^{-\frac{1}{2}\chi^2}d\chi^2 / \int_0^{\infty} \chi^{k-3}e^{-\frac{1}{2}\chi^2}d\chi^2. \quad . \quad (7)$$

Tables of this function P have been computed for various values of k, the number of classes, and χ^2.

Degrees of Freedom in χ^2. When the class probabilities p_{ij} are given *à priori* the distribution of χ^2 for k classes is expressed, as we have seen, by

$$dp = c\chi^{k-3}e^{-\frac{1}{2}\chi^2}d\chi^2. \quad . \quad . \quad . \quad (1)$$

But the presumptive class probabilities are not always given *à priori*; in a contingency table, for example, they may be *estimated* by recombining in multiplication the marginal relative frequencies of the table which is being tested. Now such a procedure *forces* the marginal totals of the presumptive table of independence to agree with those of the contingency table. This forcing reduces the number of independent class deviations from expectation. For example, in a 4-by-6 table there are 24 classes, of which 23 have independent frequencies, since the total of relative frequencies must be 1. This is in the absence of forcing. On the other hand, if the 10 marginal totals are preassigned, then there are only 3×5 or 15 independent class frequencies, as may be seen by putting these 15 in the top left part of the table, so as to fill 3 rows and 5 columns, and observing that all the others can then be filled in by reference to the marginal frequencies. In general, in an $h\times j$ table with forced marginal

agreement, there are only $(h-1)(j-1)$ independent class frequencies.

Now, in preparing for comparison by the χ^2-test in such a case, we should not integrate $c \exp(-\frac{1}{2}\chi^2) d\epsilon_1 d\epsilon_2 \ldots d\epsilon_{k-1}$ over all the previously independent ϵ_j, for by so doing we would be unfairly including combinations of the ϵ_j which have been precluded by the procedure of forced agreement. We ought to transform χ^2 so that it is expressed in terms of the *restricted* set of independent ϵ_j. It was shown by R. A. Fisher that when this is done the modified element of probability is simply

$$dp = c\chi^{k-m-3} e^{-\frac{1}{2}\chi^2} d\chi^2, \qquad . \quad . \quad (2)$$

where m is the number of restrictive relations, reducing the number of independent ϵ_j from $k-1$ to $k-m-1$. It is usual to call $k-m-1$ the number of *degrees of freedom*.

The table of $P(\chi^2)$ is therefore best constructed, and consulted, with reference not to k, the number of classes, but to $k-m-1$, the number of degrees of freedom ; and this applies not only to contingency tables but to all situations in which a presumptive probability distribution is obtained from a frequency distribution by a partial forcing of agreement, the equating of moments for example, involving restrictions on the deviations ϵ_j. These restrictions must be linear, that is to say, they must involve the ϵ_j in the 1st degree only.

Since in the deduction of $P(\chi^2)$ we excluded the case of very small class probabilities, we must exclude in practice small class frequencies. It is customary, therefore, in applying the test, to pool the small frequencies at the ends of a distribution so as to make the classes contain at least 10 individuals.

Example. The fitting of Poissonian and Type B functions to the Rutherford-Geiger data in **42**. We pool the classes corresponding to $x = 10$ and over. Thus $k = 11$.

For the Poissonian fitting there are 9 degrees of freedom, since the total frequency and the mean have been made to

agree in fitted curve and data. We find $\chi^2 = 12{\cdot}8$, and reference to tables shows that $P = 0{\cdot}20$, a satisfactory value.

For the Type B there are 8 degrees of freedom, total frequency, mean and second factorial moment having been made to agree in fitted curve and data. We find $\chi^2 = 10{\cdot}2$, $P = 0{\cdot}25$. The slight improvement is of little consequence; in both cases the principal contribution to χ^2 comes from the large deviation in class $x = 8$.

Empirical Formula. The value of χ^2 for which $P = 0{\cdot}05$ is often regarded as a boundary between the reasonable and the dubious. This value of χ^2 is given with adequate approximation, for k' degrees of freedom, by

$$1{\cdot}55(k'+2),\ k' \leqslant 10, \quad \text{and} \quad 1{\cdot}25(k'+5),\ k' > 10.$$

For $k' = 35$ the second formula above gives the value 50, the actual value of χ^2 being 49·79. For higher values of k', $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be treated as a standard normal variate.

55. Coefficients of Contingency. The possibility of dependence between variates in a contingency table can be tested by $P(\chi^2)$. For Galton's data of eye-colours in **54** the value of χ^2 is 266, a value so large that the probability of independence of eye-colour between fathers and eldest sons is negligibly small.

Attempts have been made to measure the *strength* of a dependence by means of coefficients of contingency. Thus χ^2 measures, as it were, the dispersion of a grouped sample from expectation, taken over all n individuals; and so the *mean* dispersion per individual is χ^2/n, a coefficient denoted by ϕ^2 and called by K. Pearson the *mean square contingency*. Since

$$\phi^2 = \chi^2/n = \Sigma(\epsilon_j/n)^2/p_j, \tag{1}$$

it appears that ϕ^2 is the sum of squared deviations of class *relative* frequencies ϵ_j/n from the presumptive class probabilities p_j, each divided by that probability p_j.

Pearson, considering the value of ϕ^2 for a bivariate

normal correlated distribution divided into grades of indefinite fineness in x and y, found the relation

$$\phi^2 = \rho^2(1-\rho^2), \quad \rho^2 = \phi^2/(1+\phi^2), \qquad . \quad (2)$$

and, proceeding by analogy, defined a general *coefficient of mean square contingency* C by

$$C^2 = \phi^2/(1+\phi^2). \qquad . \quad . \quad . \quad (3)$$

Evidently C^2 is zero when ϕ^2 is zero and tends to 1 as ϕ^2 increases ; but its interpretation for intermediate values is not very definite.

Example. The computation of ϕ^2 and C^2 for Galton's data in **54**.

The table of values $(p_{ij}-p_i p'_j)^2/p_i p'_j$ is :

	E_1	E_2	E_3	E_4	
F_1	0·046	0·004	0·012	0·011	0·073
F_2	0·004	0·032	0·000	0·012	0·048
F_3	0·006	0·002	0·036	0·000	0·044
F_4	0·020	0·007	0·001	0·073	0·101
	0·076	0·045	0·049	0·096	0·266

Thus $\phi^2 = 0{\cdot}266$, $C^2 = 0{\cdot}266/1{\cdot}266 = 0{\cdot}210$, $C = 0{\cdot}46$.

Table of $\mathbf{P}(\chi^2)$. A table of $P(\chi^2)$, arranged in a compact and practical form, is given in Table III of R. A. Fisher's *Statistical Methods for Research Workers*, 8th edition, pp. 110-111 ; also in the *Statistical Tables for Biological, Agricultural and Medical Research* of Fisher and Yates (Oliver and Boyd, 1938), p. 27.

For practice in the χ^2-test, the reader may examine whether the experimental data of the examples on pp. 49 and 50 are in good accord with the theoretical distributions, rectangular and binomial, there given.

CHAPTER VI

THE METHOD OF LEAST SQUARES: MULTIVARIATE CORRELATION: POLYNOMIAL AND HARMONIC REGRESSION

56. Multivariate Regression. When distributions in more than two correlated variates are encountered, an important question is the determination of the optimal value (sometimes in the sense of mean value, sometimes in the sense of most probable value) of a particular variate in terms of the values of all or any given set of the other variates. We have seen that in normal bivariate distributions the loci of such optimal values are straight regression lines. In normal correlation of many variates the corresponding loci are still linear, expressed by equations of the first degree. For three variates there are three planes of regression, for n variates there is a sheaf of n hyperplanes, each given by a linear equation expressing a particular variate in terms of the other $n-1$ variates.

It was proved by Yule that these various linear loci could be obtained without the assumption of normal distribution by using the method of Least Squares, which we now describe.

57. The Method of Least Squares. The method of Least Squares originated in the practical necessity of combining discrepant observations of a single unknown constant, or discrepant observational equations in several unknowns, in such a way as to obtain best estimates of the unknown or unknowns, under some accepted criterion.

Discrepant measures are inevitable in repeated observations, even when every effort has been made to keep conditions constant. The conditions can never be identically realized a second time. However delicate the instrument of measurement, there are innumerable fine and uncontrollable variations inherent in its parts and their adjustment and the readings, to say nothing of the inaccuracies of the observer. Hence, just as in the throwings (4) of a coin, we have varying phases of a system S. Thus repeated measures of a supposedly unique physical constant are found to be discordant, the truth being that they are a sample from a certain probability distribution depending on S. In the same way, when linear combinations or other functions of several unknowns are measured, the number of observations exceeding the number of unknowns, the equations so derived are nearly always found to be inconsistent.

In 1805 Legendre proposed, as a convenient method for reducing certain astronomical observations, that the " best value " should be taken as that for which the sum of squared deviations of the observations was least. This is the principle of *Least Squares*. It can be justified under the assumptions (i) that the measures are *normally distributed* and (ii) that the best value has *maximum probability density*. This derivation is mathematically the simplest and most rapid, but it unduly limits the types of error distribution. A more comprehensive derivation postulates that the best value is (i) a consistent or *unbiassed linear combination* of the observations and (ii) has *minimum variance*. It is remarkable that the two quite different sets of postulates lead to exactly the same equations for the unknown or unknowns.

58. Precision, Weight, Errors and Residuals. Measuring instruments of differing precision may be characterized by their standard error, or variance of error, in the reading given by them of some assigned measure.

The variance may be estimated by repeated trials. It is traditional here to use the term *weight*, defined as proportional to the reciprocal of variance of error. For example, if in determining a distance of 5000 yards the standard error of a range-finder A is estimated to be half that of a range-finder B, the weights w_A and w_B assigned to readings made by A and B would be as 4 to 1, in favour of A.

Finally, it must always be kept in mind that "true" values (if indeed the word "true" admits at all of definite meaning) are unknown and must remain unknown; so that the errors, being deviations from an unknown value, are likewise unknown. True values must be *estimated* by appropriate substitutes, namely, best or optimal values, and errors by the deviations of the observed from the optimal values. These deviations are distinguished from the errors which they represent by being called *residuals*. Errors are $\epsilon_j = a_j - a$, residuals are $e_j = a_j - \hat{a}$, where a is the true value, a_j an observed value of a, and $\hat{a}$ the optimal value of a. If there are n observations, the n residuals are estimates of the n errors; and the n errors are themselves only a finite selection under the law of probability which characterizes the circumstances of measurement.

59. Repeated Measurements of a Single Unknown. The estimate by Least Squares is found by minimizing the sum of weighted squares of residuals. The minimum of

$$S^2 = \sum_j w_j(x_j - \hat{x})^2 \tag{1}$$

is given by $\partial S^2/\partial \hat{x} = 0$, so that

$$\hat{x} = \Sigma w_j x_j / \Sigma w_j. \tag{2}$$

The optimal value of x thus appears as a *weighted mean* of the observations. If the observations are all of equal weight the optimal value is thus the arithmetic mean.

Variance of Optimal Value. The variance of $\hat{x}$ in (2) is (15)

$$\Sigma w_j^2\sigma_j^2/(\Sigma w_j)^2 = (\Sigma w_j^2\sigma^2/w_j)/(\Sigma w_j)^2 = \sigma^2/\Sigma w_i, \quad . \quad (3)$$

where σ_j^2 is the variance of x_j and σ^2 is the variance of an observation of unit weight. Thus the weight of $\hat{x}$ is Σw_j, the sum of the weights of the x_j. In particular the weight of the arithmetic mean of n values x_j is n times the weight of any x_j.

Variance of Residuals in Case of Equal Weight. If the observations are all of unit weight the j^{th} residual e_j is

$$x_j - \hat{x} = (n-1)x_j/n - (x_1+x_2+\ldots+x_n-x_j)/n. \quad . \quad (4)$$

Thus the variance of e_j is (15)

$$(n-1)^2\sigma^2/n^2 + (n-1)\sigma^2/n^2 = (n-1)\sigma^2/n. \quad . \quad (5)$$

It follows that an estimate of σ^2 is given by dividing the sum of squared residuals not by n but by $n-1$.

Ex. 1. The author made 30 bisections by eye of lines of constant length. The distribution of x, the length in cm. of the segment to the left of the point of bisection, was :

x	7·6	7·65	7·75	7·8	7·85	7·9	8·0	8·1	8·15	8·2	8·25	8·45	n
nf	2	3	1	4	4	2	4	2	3	2	2	1	30

Estimate the length of the half line and the standard error.

Ex. 2. Do the same for the results given by a second person :

x	7·7	7·75	7·8	7·85	7·9	7·95	8·0	8·05	8·1	8·15	8·2	8·3	n
nf	1	1	1	4	3	5	4	5	3	1	1	1	30

Ex. 3. Compare the precision of the two persons by assigning weights. By a weighted combination estimate the length of half the line from all 60 bisections, and assign a standard error. (The length of the line was actually 16 cm.)

60. Indirect Determinations from Linear Equations. In this case we have measurements of n linear functions of m unknowns, where n exceeds m. Because

of observational error the equations are inconsistent. For example, we might have

Observations.		Weights.
x	$= 1{\cdot}75$	2
$x+y$	$= 3{\cdot}10$	1
$x+y+z$	$= 3{\cdot}85$	4
$y+z+u$	$= 4{\cdot}30$	2
$z+u$	$= 3{\cdot}05$	3
u	$= 2{\cdot}10$	1

. . (1)

In such a case the method of Least Squares consists again in taking as optimal values those for which the sum of weighted squares of residuals is a minimum, so that for example to solve the equations (1) we would minimize

$$S^2 = 2(x-1{\cdot}75)^2+(x+y-3{\cdot}10)^2+4(x+y+z-3{\cdot}85)^2 \quad (2)$$
$$+2(y+z+u-4{\cdot}30)^2+3(z+u-3{\cdot}05)^2+(u-2{\cdot}10)^2$$

with respect to x, y, z, u. More generally, if the equations are (to take the case of 4 unknowns)

$$\begin{aligned} a_1x+b_1y+c_1z+d_1u &= h_1, \text{ weight } w_1, \\ a_2x+b_2y+c_2z+d_2u &= h_2, \ \ldots \ w_2, \\ &\cdots \\ a_mx+b_my+c_mz+d_mu &= h_n \qquad w_n \end{aligned} \quad (3)$$

we minimize

$$S^2 = \sum_{j=1}^{m} w_j(a_jx+b_jy+c_jz+d_ju-h_j)^2, \quad (4)$$

and similarly for any number of unknowns.

The partial derivatives $\partial S^2/\partial x$, $\partial S^2/\partial y$, $\partial S^2/\partial z$, $\partial S^2/\partial u$ must be zero ; and so we derive the equations

$$\begin{aligned} (\Sigma w_ja_j^2)x &+(\Sigma w_ja_jb_j)y+(\Sigma w_ja_jc_j)z+(\Sigma w_ja_jd_j)u=\Sigma w_ja_jh_j, \\ (\Sigma w_ja_jb_j)x&+(\Sigma w_jb_j^2)y &+(\Sigma w_jb_jc_j)z+(\Sigma w_jb_jd_j)u=\Sigma w_jb_jh_j, \\ (\Sigma w_ja_jc_j)x&+(\Sigma w_jb_jc_j)y+(\Sigma w_jc_j^2)z &+(\Sigma w_jc_jd_j)u=\Sigma w_jc_jh_j, \\ (\Sigma w_ja_jd_j)x&+(\Sigma w_jb_jd_j)y+(\Sigma w_jc_jd_j)z+(\Sigma w_jd_j^2)u &=\Sigma w_jd_jh_j, \end{aligned} \quad (5)$$

for x, y, z, u. These are called the *normal* equations, and their general form is similar to the above. Inspection will

show that the coefficients in the normal equations are *symmetrical*, in the respect that the coefficient of the j^{th} unknown in the k^{th} equation is identical with that of the k^{th} unknown in the j^{th} equation. The scheme of coefficients is in fact symmetrical about its northwest to southeast diagonal. This symmetry is of great service in shortening the solution of the equations.

Thus in our numerical example the normal equations will be found to be

$$\begin{aligned} 7x+5y+4z &= 22{\cdot}000, \\ 5x+7y+6z+2u &= 27{\cdot}100, \\ 4x+6y+9z+5u &= 33{\cdot}150, \\ 2y+5z+6u &= 19{\cdot}850, \end{aligned} \qquad . \quad . \quad (6)$$

which can now be solved by methods of practical algebra. The solutions are $x = 1{\cdot}750$, $y = 1{\cdot}274$, $z = 0{\cdot}846$, $u = 2{\cdot}178$.

Various schemes of systematic solution of normal equations have been devised, and for these the reader must be referred to more comprehensive treatises and original memoirs dealing with Least Squares or with the numerical solution of algebraic equations.

Preparation of Normal Equations. It is evident from the construction of the sum S^2 of weighted and squared residuals that exactly the same sum would arise if we multiplied each observation throughout by the square root of its weight, $\sqrt{w_j}$, and then regarded the observational equations as of equal unit weight. (Let the reader verify this from the example.) Such a reduction of a set of equations with unequal weights to a set with equal weights is called *preparing* the equations.

61. Application of Least Squares to Trivariate Correlation. Suppose that we have n trivariate observations (x_j, y_j, z_j), as for example the height, weight and chest measurement of each of 1000 soldiers, and that we wish to express each variate as the best possible linear estimate of the other two. We may suppose the variates

measured as deviations from their respective means, and standardized. Thus for x we have n equations

$$x_j = b_{12}y_j + b_{13}z_j, \ (j = 1, 2, 3, \ldots, n). \tag{1}$$

These may be regarded as n observational equations in the two unknowns b_{12} and b_{13}. If we solve them by least squares we shall have the desired optimal relation

$$\hat{x} = b_{12}y + b_{13}z, \tag{2}$$

which may be regarded geometrically as the *regression plane* of x on y and z. The coefficients b_{12} and b_{13} are called regression coefficients ; they are the sample estimates of ideal regression coefficients β_{12}, β_{13} in the underlying population. The normal equations for b_{12} and b_{13} are obtained by minimizing the sum of squared residuals

$$S^2 = \sum_j (x_j - b_{12}y_j - b_{13}z_j)^2. \tag{3}$$

The minimum conditions $\partial S^2/\partial b_{12} = 0$, $\partial S^2/\partial b_{13} = 0$ give, on division by n,

$$\begin{aligned} b_{12} + r_{23}b_{13} &= r_{12}, \\ r_{23}b_{12} + b_{13} &= r_{13}, \end{aligned} \tag{4}$$

where $r_{12} = \Sigma x_j y_j/n$, $r_{13} = \Sigma x_j z_j/n$, $r_{23} = \Sigma y_j z_j/n$.

Solving, we find the desired regression coefficients as

$$\begin{aligned} b_{12} &= (r_{12} - r_{13}r_{23})/(1 - r_{23}^2), \\ b_{13} &= (r_{13} - r_{12}r_{23})/(1 - r_{23}^2), \end{aligned} \tag{5}$$

and similar results hold for the regression of y on x and z, and of z on x and y.

The standardized mean-product-deviations r_{12}, r_{13} and r_{23} are usually called *total correlation coefficients* of x and y, x and z and y and z respectively. They are really estimates from sample of the corresponding mean-product-deviations, or product-moments ρ_{12}, ρ_{13} and ρ_{23} in the trivariate population or probability function.

It may be proved that the trivariate normal m.g.f., in standard scale and with means as origin, is

$$\exp(\alpha^2+\beta^2+\gamma^2+2\rho_{12}\alpha\beta+2\rho_{13}\alpha\gamma+2\rho_{23}\beta\gamma) \quad . \quad (6)$$

and by reversion that the corresponding trivariate normal function is

$$\phi(x, y, z)$$
$$= (2\pi)^{-\frac{3}{2}}\triangle^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}\triangle^{-1}\{(1-\rho_{23}^2)x^2+(1-\rho_{13}^2)y^2+(1-\rho_{12}^2)z^2\right.$$
$$\left.-2(\rho_{12}-\rho_{13}\rho_{23})xy-2(\rho_{13}-\rho_{12}\rho_{23})xz-2(\rho_{23}-\rho_{12}\rho_{13})yz\}\right], \quad (7)$$

where $\triangle$ is the determinant

$$\triangle = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{vmatrix}$$

of total correlations.

The equations $\partial\phi/\partial x = 0$, $\partial\phi/\partial y = 0$, $\partial\phi/\partial z = 0$ give the loci of maximum probability of x for fixed y and z, of y for fixed x and z, and of z for fixed x and y. By actual differentiation we find these loci to be

$$x = \beta_{12}y+\beta_{13}z \quad . \quad . \quad . \quad (8)$$

and two others, where

$$\begin{aligned} \beta_{12} &= (\rho_{12}-\rho_{13}\rho_{23})/(1-\rho_{23}^2), \\ \beta_{13} &= (\rho_{13}-\rho_{12}\rho_{23})/(1-\rho_{23}^2). \end{aligned} \quad (9)$$

Thus we see that the estimates of regression by Least Squares are in agreement with those based on normal trivariate correlation. A corresponding result is true for linear regression in any number of variates.

62. Partial Correlation. The unstandardized equations, with means as origin, of the regression lines in bivariate regression (49) are

$$\begin{aligned} x &= \beta_{12}y, \text{ where } \beta_{12} = \rho\sigma_1/\sigma_2, \\ y &= \beta_{21}x, \text{ where } \beta_{21} = \rho\sigma_2/\sigma_1. \end{aligned} \quad . \quad . \quad (1)$$

The correlation coefficient ρ appears here as the geometric mean $(\beta_{12}\beta_{21})^{\frac{1}{2}}$. On the analogy of this, partial

correlation coefficients in multivariate problems have been defined as the geometric means of the corresponding regression coefficients. For example, the partial coefficient of two variables x_h and x_k would be defined by $(\beta_{hk}\beta_{kh})^{\frac{1}{2}}$.

Notation. It is customary to denote, for example in a four-variate problem, the partial correlation coefficient of x and y by $\rho_{12,\,34}$, to distinguish it from the total correlation coefficient ρ_{12}. The sample estimate would be written $r_{12,\,34}$.

Example. Given the following estimates of variances and total correlations of three variables x, y, z, find the three regression equations and the three estimates of partial correlation coefficients:

$$\sigma_1^2 = 5{\cdot}0,\ \sigma_2^2 = 7{\cdot}0,\ \sigma_3^2 = 3{\cdot}0,\ r_{12} = 0{\cdot}80,\ r_{13} = 0{\cdot}40,$$
$$r_{23} = 0{\cdot}60.$$

63. Non-Linear Regression : Polynomial Regression. From the nature of a set of observations of a variate y dependent on x it may be apparent that the regression cannot be linear. Common types of non-linear regression are those in which the underlying functional relation of y and x is of polynomial, or of harmonic type.

The polynomial regression

$$y = c_0 + c_1x + c_2x^2 + \ldots + c_kx^k \qquad . \quad . \quad (1)$$

will be considered first in its simplest case, the fitting of the polynomial by Least Squares to n independent observations u_x of equal weight, corresponding to n equispaced values of x, namely $x = 0, 1, \ldots, n-1$.

The polynomial of best fit is given by the minimum of the sum of squared residuals

$$S^2 = \sum_x (u - c_0 - c_1x - c_2x^2 - \ldots - c_kx^k)^2, \qquad . \quad (2)$$

that is, by the conditions $\partial S^2/\partial c_j = 0$. These give $k+1$ normal equations for the c_j, easily seen to be expressible as

$$\Sigma x^j(u_x - y_x) = 0, \quad (j = 0, 1, 2, \ldots, k), \qquad . \quad (3)$$

displaying the fact that the fitting of a polynomial of degree k by Least Squares is equivalent to equating the moments of orders 0, 1, 2, ..., k of the polynomial and the data.

The values of the coefficients c_j can be found by solving the normal equations, but since sums of powers of natural numbers up to the $2k^{th}$ are required, the method becomes laborious if n is large and if the polynomial y is of the 3rd or higher degree. For this reason it is better to express y not in powers of x, but in polynomials 1, $t_1(x)$, $t_2(x)$, ..., $t_k(x)$ having the property of being *uncorrelated*, being in fact such that the product sum

$$\sum_x t_i(x)t_j(x) = 0 \text{ if } i \neq j. \quad . \quad . \quad . \quad (4)$$

These polynomials $t_j(x)$ are familiar in mathematics as the *orthogonal* polynomials of Tchebychef, and their properties are known. For example, it is known that

$$t_j(x) = (2j)_{(j)}x_{(j)} - (2j-1)_{(j)}(n-j)x_{(j-1)} + (2j-2)_{(j)}(n-j+1)_{(2)}x_{(j-2)} - \ldots + (-)^j(n-1)_{(j)}, \quad (5)$$

so that (Appendix, **1**) the s^{th} difference

$$\Delta^s t_r(x) = (2j)_{(j)}x_{(j-s)} - (2j-1)_{(j)}(n-j)x_{(j-s-1)} + \ldots + (-)^{j+s}(j+s)_{(j)}(n-s-1)_{(j-s)}. \quad (6)$$

It is also known that

$$\sum_x (t_j(x))^2 = n(n^2-1)(n^2-4)\ldots(n^2-j^2)/[(2j+1)(j!)^2]. \quad (7)$$

If, therefore, we express y in the form

$$y = a_0 + a_1t_1(x) + a_2t_2(x) + \ldots + a_kt_k(x), \quad . \quad . \quad (8)$$

the sum S^2 of squared residuals, because of the vanishing of the product terms, takes the form

$$S^2 = \sum [u_x - a_0 - a_1t_1(x) - \ldots - a_kt_k(x)]^2$$

$$= \sum_x [(u_x)^2 - 2u_x(a_0 + a_1t_1(x) + \ldots + a_kt_k(x)) + a_0^2 + \ldots + (a_kt_k(x))^2] \quad (9)$$

and the normal equations $\partial S^2/\partial a_j = 0$ therefore take the form

$$a_j = \Sigma u_x t_j(x)/\Sigma(t_j(x))^2. \quad (j = 0, 1, \ldots, k). \quad . \quad (10)$$

Thus each coefficient a_j in the regression is found independently of the others, without the labour of solving simultaneous equations. (The choice of uncorrelated or orthogonal functions for the representation of u_x always confers this very great advantage.) Since the polynomials $t_j(x)$ are expressed in factorials $x_{(r)}$, the numerator of the expression for a_j can easily be found in terms of the *factorial moments* of the data u_x, these moments being obtained as usual (Appendix, 2) by summation.

The minimum sum of squared residuals can itself be evaluated beforehand, for by (9) and (10) it takes the form

$$\begin{aligned}&\Sigma[u_x^2 - a_0^2 - a_1^2(t_1(x))^2 - \ldots - a_k^2(t_k(x))^2] \\ &= \Sigma u_x^2 - a_0\Sigma u_x - a_1\Sigma u_x t_1(x) - \ldots - a_k\Sigma u_x t_k(x), \quad (11)\end{aligned}$$

involving the sum of the squares of the u_x, diminished by the product of each successive a_j by the numerator in (10). It is known that the variance of a single residual is best estimated by dividing the sum of the n squared residuals by the degrees of freedom, $n-k-1$; hence we can judge beforehand, if we know the precision of the data, what value of k gives the best polynomial y. It is of course possible, by taking too many terms in the polynomial y, to fit the data too well, in the sense that the sum of squared residuals is much smaller than that warranted by the precision of the data.

64. Practical Routine of Fitting a Polynomial. All of the above points, which can be treated only briefly here, have been discussed at length in special memoirs. We shall merely illustrate a method depending on the theory of 63 and making use of a table containing the terminal values and differences $t_j(0)$, $\Delta t_j(0)$, $\Delta^2 t_j(0)$, ..., for $j = 0, 1, 2, 3, \ldots, k$ and the particular value of n.

The rule for constructing such a table follows from (5) and (6) and is simple. We shall illustrate it for $n = 6$. We write down the fixed table of binomial coefficients, table (i) below, to $k+1$ columns ; in the illustration, $k = 3$. Beside table (i) we place table (ii), consisting of binomial coefficients of $n-1$, $n-2$, ... written below each other as shown, also to $k+1$ columns. The products of corresponding entries in the two tables now give us the desired table (iii) of terminal values and differences of t-polynomials, and at the feet of the respective columns we enter the values of Σt_j^2, as computed from the formula 63 (7).

(i)

1	−1	1	−1
	2	−3	4
		6	−10
			20

(ii)

1	5	10	10
	1	4	6
		1	3
			1

iii)

1	−5	10	−10
	2	−12	24
		6	−30
			20
6	70	336	720

(iv)

1	−5	5	− 5
	2	−6	12
		3	−15
			10
6	70	84	180

A possibility making table (iii) still simpler for practical use is that when a common integer factor is observed in any column, we may cancel through by that factor, provided that the square of that factor is cancelled through from Σt_j^2. Thus the cancelling of factors 2, 2 from columns 3, 4 in table (iii) above gives table (iv). Such tables, extended to six or seven columns, are easily constructed for a proposed value of n.

The use of the table in finding the regression coefficients a_j and the fitted values y_x is best illustrated by an actual worked example. The process is no more difficult for a long series of data than for a short, but to economize in space we shall illustrate it by fitting a cubic polynomial to six values u_x.

Example.

0	1	2	3	4	5
5	13	25	60	105	200

By summation the reduced factorial moments of u are found to be 408, 1663, 2835 and 2480, while $\Sigma u^2 = 55444$.

Using four columns only of the table of polynomials (since we are fitting a cubic) we set out the rest of the work in compact shape thus :

	408	1286	567	191				
a_j	68	18·371	6·75	1·0611		Δy_0	$\Delta^2 y_0$	$\Delta^3 y_0$
Sums	1	−5	5	−5	4·590			
408		2	−6	12		8·975		
1663			3	−15			4·334	
2835				10				10·611
2480								
	6	70	84	180	Check $y_5 = 198·91$.			

Explanation.

$a_0 = (408 \times 1)/6 = 68.$
$a_1 = (1663 \times 2 - 408 \times 5)/70 = 1286/70 = 18·371.$
$a_2 = (2835 \times 3 - 1663 \times 6 + 408 \times 5)/84 = 567/84 = 6·75,$

and so on; the elements in *columns* of the table are used as multipliers of the factorial moments, the entries at the feet of the columns as divisors. Then

$$\begin{aligned} y_0 &= 68 \times 1 - 18·371 \times 5 + 6·75 \times 5 - 1·0611 \times 5 = 4·590. \\ \Delta y_0 &= \quad 18·371 \times 2 - 6·75 \times 6 + 1·0611 \times 12 = 8·975. \\ \Delta^2 y_0 &= \quad\quad 6·75 \times 3 - 1·0611 \times 15 = 4·334, \end{aligned}$$

and so on; the elements in *rows* are now used as multipliers of the a_j, and give the terminal value y_0 and its differences. There is also a good check on the other terminal value,

$$y_5 = 68 \times 1 + 18·371 \times 5 + 6·75 \times 5 + 1·0611 \times 5 = 198·91,$$

the same terms as gave y_0, but with positive multipliers.

Building up a difference table of the y_x from the constant 3rd differences in the way familiar in interpolation, we have—

x	y	$\triangle y$	$\triangle^2 y$	$\triangle^3 y$	u	$u-y$
0	4·590				5	0·410
		8·975				
1	13·565		4·334		13	−0·565
		13·309		10·611		
2	26·874		14·945		25	−1·874
		28·254		10·611		
3	55·128		25·556		60	4·872
		53·810		10·611		
4	108·938		36·167		105	−3·938
		89·977				
5	198·915				200	1·085

The comparison of the fitted values with the data can be seen in the columns headed y and u. The sum of squared residuals $(u-y)^2$ will be found to be 44·4.

But we can also set out a table thus, estimating by **63** (11) the variance of a residual after a constant, a straight line, a parabola and our cubic are fitted in succession :

k	$n-k-1$	a_k	num. of a_k	prod.	S^2	$\div(n-k-1)$
					55444	
0	5	68	408	27744	27700	5540
1	4	18·371	1286	23625	4075	1019
2	3	6·75	567	3827	248	83
3	2	1·0611	191	203	45	22

The column headed S^2 shows the sum of squared residuals, obtained in accordance with **63** (11) by subtracting the entries in the previous column in turn from $\Sigma u^2 = 55444$. The last column gives estimates of the variance of a single residual at the different stages. To test which polynomial best represents the data, we must have a preliminary knowledge or estimate of the variance of the observations. This variance is compared with the residual variance in the light of the sampling distributions of **71** and **74**.

The alternative computation of the sum of squared residuals as 45 checks the work, for the same sum was given by the fitted values as 44·4.

For a given value of n the *same* table of terminal values and differences of t-polynomials serves for fitting a poly-

nomial of any degree. Thus, using only the first three columns of the table in the above worked example, we may fit a parabola instead of a cubic. It will be found instructive to do this, following the details of the worked example. Notice that the coefficients a_0, a_1, a_2 are the same as before.

Example. Fit a cubic polynomial to the seven equidistant and equally weighted data

x	0	1	2	3	4	5	6
u	−11	5	13	25	60	105	200

65. Periodic Regressions : Observations of Equal Weight. Observations which exhibit periodicity more or less masked by accidental error are of common occurrence. The height of tide-water at a seaport, measured at equal intervals of time, shows such a periodicity; monthly averages of temperature show a seasonal periodicity; telephone calls on an Exchange show a weekly periodicity.

The procedure for analysing periodicity is to assume a periodic function

$$y_\theta = a_0 \begin{matrix} +a_1 \cos\theta + a_2 \cos 2\theta + \ldots + a_k \cos k\theta \\ +b_1 \sin\theta + b_2 \sin 2\theta + \ldots + b_k \sin k\theta \end{matrix} \qquad (1)$$

and to find the coefficients a_j and b_j of the constituent periodic terms by the method of Least Squares.

We consider therefore n equally spaced observations u_θ of equal weight, where $\theta = 0, 2\pi/n, 4\pi/n, \ldots, 2(n-1)\pi/n$, the observations thus corresponding to the n phase-angles of one complete oscillation of a periodic phenomenon. The initial observation of a second oscillation is not included. In view of the trigonometrical relations

$$\sum_{r=0}^{n-1} \cos\frac{2rh\pi}{n} \cos\frac{2rj\pi}{n} = \begin{cases} 0, \text{ if } h \neq j, \\ \frac{1}{2}n, \text{ if } h = j \neq 0, \neq \frac{1}{2}n, \end{cases} \qquad (2)$$

and the similar ones with one or both cosines replaced by sines (these are really *orthogonal* relations exactly

resembling those of the Tchebychef polynomials in 63 (4) and (7)), the sum S^2 of squared residuals is (*cf.* 63 (9))

$$\sum_\theta (u_\theta - y_\theta)^2 = \sum_\theta [u_\theta^2 - 2u_\theta(a_0 + a_1 \cos\theta + \ldots + b_k \sin k\theta)] + \tfrac{1}{2}n(2a_0^2 + a_1^2 + \ldots + b_k^2). \quad . \quad (3)$$

Differentiating with respect to the a_j and b_j and equating to zero, we have the normal equations for the regression coefficients. Each is given independently of the others.

$$a_0 = \frac{1}{n}\sum_\theta u_\theta, \quad a_h = \frac{2}{n}\sum_\theta u_\theta \cos h\theta, \quad b_h = \frac{2}{n}\sum_\theta u_\theta \sin h\theta. \quad . \quad (4)$$

If n is even,

$$a_{\frac{1}{2}n} = \frac{1}{n}\sum_\theta u_\theta \cos \tfrac{1}{2}n\theta, \quad b_{\frac{1}{2}n} = 0, \qquad (5)$$

and $\cos \frac{1}{2}n\theta$ is $+1$ and -1 alternately as θ takes its n values.

The theoretical solution is thus immediate. Simplicity of practical application will depend on the value of n, and the consequent values of $\cos h\theta$ and $\sin h\theta$.

66. Practical Solution of the Normal Equations. The process of numerical solution becomes specially simple when θ, that is, $2\pi/n$, is such that $\cos h\theta$ and $\sin h\theta$ are easy to handle. This occurs when $n = 4$, 6, 8, 12 or 24, the last two cases being specially important, as corresponding to the hourly or two-hourly subdivision of the day; and special routines for these values of n have been devised.

The procedure depends on the fact that in the four quadrants, from $\theta = 0$ to $\theta = 2\pi$, $\cos\theta$ and $\sin\theta$ take the same absolute values four times, though with differing alternations of sign. To take the case $n = 12$ for illustration, the data u_θ (and there will be no misunderstanding if these are written meanwhile as u_0, u_1, ..., u_{11}) can be assembled in tetrads, for example $u_1 + u_5 - u_7 - u_{11}$, before

being multiplied by the suitable values of $\cos h\theta$ and $\sin h\theta$, where $h\theta$ can always be taken as coterminous with some angle in the first quadrant.

We shall indicate how this is done by an actual example.

Example. To fit terms as far as $a_4\cos 4\theta$, $b_4\sin 4\theta$ to the 12 data (Whittaker and Robinson, *Calculus of Observations*, p. 272):

u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}
2·71	3·04	2·13	1·27	0·79	0·50	0·37	0·54	0·19	−0·35	−0·44	0·77

First write the data in a scheme U of columns, down, up, down, up, with blanks as indicated by the dots, as follows:

U	271	37	.		M	+ + + +
	304	50	54	77		+ − − +
	213	79	19	-44		+ − + −
	127	.	−35			+ + − −

Next, add along the rows of the scheme U, after giving sign to the columns of U in four different ways, according to the rows in the sign-scheme M. We thus obtain four separate sets of totals, and these are combined with cosines and sines of 0°, 30°, 60°, 90° in four separate schemes, as below. (We have included the coefficients necessary for computing a_5, b_5 and a_6 as well.)

		a_2		a_6		a_1	a_3	a_5
308	0·5	1	1	0·5	234	1	1	1
485	0·5	0·5	−0·5	−0·5	277	0·866	0	−0·866
267	0·5	−0·5	−0·5	0·5	71	0·5	−1	0·5
92	0·5	−1	1	−0·5	162	0	0	0
6	576	325	24	−1	6	509·4	163	26·6

		b_2	b_4		b_1	b_3	b_5
234	0	0	308	0	0	0
231	0·866	0·866	223	0·5	1	0·5
197	0·866	−0·866	317	0·866	0	−0·866
92	0	0	162	1	−1	1
6	370·6	29·4	6	548	61	−0·5
	0·618	0·049		0·913	0·102	−0·001

Explanation.

$$a_1 = (2{\cdot}34\times 1+2{\cdot}77\times 0{\cdot}866+0{\cdot}71\times 0{\cdot}5+1{\cdot}62\times 0)/6$$
$$= 0{\cdot}849, \text{ etc.}$$

Hence, as far as terms in $\cos 4\theta$, $\sin 4\theta$, the regression is

$$y_\theta = 0{\cdot}960 \begin{matrix} +0{\cdot}849\cos\theta+0{\cdot}542\cos 2\theta+0{\cdot}272\cos 3\theta \\ +0{\cdot}040\cos 4\theta \\ +0{\cdot}913\sin\theta+0{\cdot}618\sin 2\theta+0{\cdot}102\sin 3\theta \\ +0{\cdot}049\sin 4\theta, \end{matrix}$$

and the regressions to fewer or more terms involve the same coefficients a_j, b_j as are given by the above scheme of solution.

The sum of squared residuals may also be calculated beforehand from the regression coefficients in a scheme set out as follows :

k	$n-2k-1$	na_0^2 and $\frac{1}{2}n(a_k^2+b_k^2)$	S^2	$\div(n-2k-1)$
			24·983	
0	11	11·059	13·924	1·266
1	9	9·326	4·598	0·511
2	7	4·054	4·544	0·078
3	5	0·506	0·038	0·008
4	3	0·024	0·014	0·005

Just as in polynomial regression, the contributions to the sum of squared residuals produced by successive terms are subtracted in turn from u^2, which here is 24·983. The estimate of variance of a single residual is then made by dividing the residual sum of squares by $n-2k-1$, the number of degrees of freedom. The results are shown in the last column.

67. General Regressions. After what has preceded, the routine to be adopted in other regressions, such as

$$y = a_0+a_1\tan\theta+a_2\tan 2\theta+\ldots+a_k\tan k\theta \quad . \quad (1)$$

will be readily understood. Such regressions are not common in statistical work, but they are not outside the bounds of possibility. The desirable thing in any problem

of regression will be to express y, if possible, in terms of functions which, like the Tchebychef polynomials or the sine and cosine of multiples of $2\pi/n$, have the orthogonal property that the product-sums of different functions of the set over the range vanish. The effective meaning of this is that the contributions of the successive terms to the regression are uncorrelated with each other.

Harmonic Analysis. For fuller details concerning the practical routine of estimating periodic regressions, the reader may consult the chapters on harmonic analysis in Whittaker and Robinson's *Calculus of Observations*, or Brunt's *Combination of Observations*, 2nd edition, 1931.

CHAPTER VII

PROBABILITY DISTRIBUTIONS OF STATISTICAL COEFFICIENTS

68. Sampling Distributions. A statistical coefficient computed from a sample of n values, univariate or multivariate, is only an estimate of the corresponding parameter in the population or underlying probability function. It is therefore to be presumed erroneous, though the degree of error cannot be affirmed exactly, since the true value of the parameter is not known. The degree of error can be stated only in terms of probability; and the probability distributions involved are (i) the hypothetical population, or distribution of the variate or variates, (ii) the derived distribution of the coefficient of estimate from sample. The second of these is called the *sampling distribution* of the coefficient.

Let us consider a case in which the first of these distributions, the probability distribution of the variate, is not hypothetical but given. In Charlier's experiment (**22**) of drawing 10 cards from a pack, with replacement of each card, and continuing this until a sample of 1000 sets of 10 cards had been collected, the variate was the number x of black cards in a set of 10, and its probability distribution was the binomial distribution, with mean 5 and variance 2·5; the corresponding values of mean and mean square deviation in Charlier's sample were 4·933 and 2·415. Are the respective deviations 4·933−5, or −0·067, and 2·415−2·5, or −0·085, reasonable or abnormal? Such questions can be answered only when

the sampling distributions of the estimates of mean and variance are known.

The nature and genesis of these sampling distributions can be illustrated from this same example. The sample group of 1000 sets of 10 card drawings was merely one out of an enormous number of equally possible groups. From the pack of 52 cards the 10 cards, drawn one at a time with replacement, could eventuate, if order of drawing were taken into account, in 52^{10} ways. This is an unimaginably large number, but the number of groups of 1000 sets which may be chosen from these 52^{10} sets is incomparably greater still. Each group may be supposed to have its mean m and mean square deviation s^2, computable in the usual way. The aggregates of these values of m and s^2 constitute probability distributions, and these are the *sampling distributions* of m and s^2 for the kind of sample in question.

Example. If the parent population is normal and the number in sample is n, the sampling variances of the estimates m_2, m_3, m_4 of the moments μ_2, μ_3, μ_4 are respectively $2\sigma^4/n$, $6\sigma^6/n$, $96\sigma^8/n$. For μ_5, μ_6, ... they increase rapidly.

The functional form of a sampling distribution depends (i) on the population (probability function of the variate or variates sampled), (ii) on the function used for estimating the parameter, and (iii) on n, the number of observations in the sample. Since 1900, and especially since 1915, much research has been expended on the problem of deriving the probability distributions of the commoner coefficients. Most of this research has been devoted to samples of a *normally* distributed variate or variates, and the sampling distributions are now well known and already classic. It appears that as the number n in sample increases the sampling distributions of many coefficients, though by no means of all, tend themselves towards the normal type. In such cases it is customary to supply an estimate of the precision of a coefficient by appending to

its computed value its standard deviation of sampling, or standard error; and this is sometimes said to imply a probability of 10/20 that the true value lies in the range delimited by twice the standard error on either side of the computed value.

The form of statement is not strictly accurate; if, for example, a computed mean is m, and the central 95 per cent. range of sampling probability area is the criterion of what is acceptable, then m may be anywhere from the extreme left of the 95 per cent. range of a sampling distribution centred on a hypothetical mean μ' to the extreme right of the 95 per cent. range of a second sampling distribution centred on a hypothetical mean μ''; but these are different distributions, and it does not follow either that the left half-range $\mu'-m$ of the first is equal to the right half-range $m-\mu''$ of the second, or that we can add the probabilities, under the different hypotheses, that the true μ lies in these respective half-ranges. In fact, μ being an *unknown* parameter, an ordinary direct statement of probability cannot be made.

When the number in sample n is small the sampling distribution of the coefficient is often of non-normal, skew or platykurtic type, and the standard error is an insufficient indication of the interval within which the true value of the parameter may lie. It is necessary in such a case to know the sampling distribution and probability integral of the special coefficient.

69. The Sampling Distribution of Means. In a few cases the sampling probability function of the mean of n observations is of the same type as the probability function of the population. For example, the normal probability function with mean μ and variance σ^2,

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \tag{1}$$

has m.g.f. $\exp(\mu a+\frac{1}{2}\sigma^2 a^2)$. Hence the m.g.f. of the sum of n sample values x_j is $\exp(n\mu a+\frac{1}{2}n\sigma^2 a^2)$. To change

from sum to mean is to write x/n for x, or a/n for a. Hence the m.g.f. of the mean of sample is $\exp(\mu a+\frac{1}{2}\sigma^2a^2/n)$. The mean of sample is thus distributed normally about the same mean as before, but with variance σ^2/n, or standard error $\sigma/\sqrt{n}$.

Ex. 1. Prove that if λ_1, λ_2, λ_3, ... are the seminvariants of any population, the seminvariants of the mean of a sample of n are λ_1, λ_2/n, λ_3/n^2, ...

Ex. 2. The number x of black cards in a set of 10 in Charlier's experiment is binomially distributed with mean 5 and variance 2·5. The mean of x in 1000 sets is distributed with approximate normality, about mean 5, and with variance 2·5/1000, or 0·0025. The standard error is thus 0·05. The deviation of the mean 4·933 of Charlier's sample from 5 is −0·067, about 4/3 of the standard error.

The deviation is not excessive. From the table of the normal probability integral on p. 144 it is seen that the probability of a deviation exceeding $1{\cdot}34\sigma$ is about 0·18.

Again, if the probability function of x is of Gamma or so-called χ^2 type, namely

$$\phi(x) = (\Gamma(k))^{-1}x^{k-1}e \quad . \quad . \quad (2)$$

the m.g.f. is

$$(\Gamma(k))^{-1}\int_0^\infty x^{k-1}e^{-x}e^{ax}dx = (1-a)^{-k}. \quad (3)$$

The m.g.f. of the sum of n sample values x_j is $(1-a)^{-nk}$, and so the m.g.f. of m, the sample mean, is $(1-a/n)^{-nk}$. Reverting to the probability function, which by a theorem of Lerch is unique, we obtain the probability function of m as

$$\phi(m) = n(\Gamma(nk))^{-1}(nm)^{nk-1}e^{-nm}. \quad . \quad . \quad (4)$$

This is again of Pearson's Type III.

Ex. 3. Prove that the distribution of the sum (not the mean) of n values x_j each obeying the Poissonian law $\psi(x)$ of 33 is Poissonian. (Use the f.m.g.f. of x.)

70. Distribution of Mean Square in Normal Sample. If the probability differential of x is

$$dp = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx, \quad . \quad . \quad (1)$$

that of $z = \frac{1}{2}x^2$, as in **37** (2), is

$$dp = \frac{1}{\sqrt{\pi}} z^{-\frac{1}{2}} e^{-z} dz, \quad (2)$$

and so the m.g.f. of z is $(1-\alpha)^{-\frac{1}{2}}$, by **69** (3). Hence the m.g.f. of half the sum of the squares of n sample values x_i is $(1-\alpha)^{-\frac{1}{2}n}$, and so if s^2 is the mean of the squares the m.g.f. of $\frac{1}{2}s^2$ is $(1-\alpha/n)^{-\frac{1}{2}n}$. It follows that the probability differential of u, where $u = \frac{1}{2}s^2$, is

$$dp = [\Gamma(\tfrac{1}{2}n)]^{-1} n^{\frac{1}{2}n} (u)^{\frac{1}{2}n-1} e^{-nu} du, \quad . \quad . \quad (3)$$

again of χ^2 type. By changing from $\frac{1}{2}s^2$ to s^2 we have the probability function of s^2, namely

$$\phi(s^2) = 2^{-\frac{1}{2}n} [\Gamma(\tfrac{1}{2}n)]^{-1} n^{\frac{1}{2}n} (s^2)^{\frac{1}{2}(n-2)} e^{-\frac{1}{2}ns^2}. \quad . \quad (4)$$

In unstandardized units we must write s^2/σ^2 for s^2 on the right of (4), and insert the factor $1/\sigma^2$.

The seminvariant g.f. of s^2 is

$$\begin{aligned} -\tfrac{1}{2}n \log (1-2\alpha/n) &= \tfrac{1}{2}n(2\alpha/n + 4\alpha^2/2n^2 + \ldots) \\ &= \alpha + 2\alpha^2/2!n + \ldots \quad . \quad . \quad . \quad (5) \end{aligned}$$

Thus the mean of s^2 is 1 and the variance is $2/n$; in unstandardized units these are σ^2 and $2\sigma^4/n$, where σ^2 is the variance of x. The s.g.f. also shows that as n increases the m.g.f. of s^2 tends to asymptotic equivalence with $\exp(\alpha + \alpha^2/n)$; and so the distribution of s^2 tends to normality.

Example. The distribution of s^2 in Charlier's 1000 sets is almost normal; $\sigma^2 = 2{\cdot}5$, and s^2 computed from the sample (using deviations not from Charlier's mean $m = 4{\cdot}933$, but from $\mu = 5$) is $2{\cdot}419$. The standard error of s^2 is $\sigma^2\sqrt{(2/n)} = 2{\cdot}5/\sqrt{500} = 0{\cdot}112$. The actual deviation, $-0{\cdot}081$, is numerically about three quarters of this.

71. Distribution of Estimate of Variance. The variance or second moment of the population is commonly estimated from the sample by taking the n^{th} part of the sum of squared deviations of the sample values x_j from the sample mean m.

Of the n deviations from m only $n-1$ are independent, and this estimate of σ^2, which we shall call s^2 though pointing out that it is not the same s^2 as in **70**, can be expressed as a quadratic expression in $n-1$ independent values. Thus we have, by **14** (5),

$$\begin{aligned}&(x_1^2+x_2^2+\ldots+x_n^2)/n-(x_1+x_2+\ldots+x_n)^2/n^2\\ &=(n-1)(z_1^2+z_2^2+\ldots+z_{n-1}^2)/n^2\\ &\quad -2(z_1z_2+z_1z_3+\ldots+z_{n-2}z_{n-1})/n^2, \quad . \qquad (1)\end{aligned}$$

where $z_1 = x_1-x_n$, $z_2 = x_2-x_n$, ..., $z_{n-1} = x_{n-1}-x_n$; and this is but one of many ways in which s^2 may be expressed in terms of only $n-1$ variables. The z_j here are linearly independent, though correlated by possessing the term $-x_n$ in common. (See Appendix, **5**.)

This loss of a degree of freedom, for that is what it is, complicates the problem of finding the distribution of s^2, but its m.g.f. can be evaluated as a multiple integral over the n sample values, and proves to be $(1-2\alpha/n)^{-\frac{1}{2}(n-1)}$, which differs from that of the s^2 in **70** only in the exponent, $n-1$ replacing n. It follows that the distribution of s^2 is again of χ^2 type, its probability function being in fact

$$\phi(s^2) = 2^{-\frac{1}{2}(n-1)}\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{-1} n^{\frac{1}{2}(n-1)}(s^2)^{\frac{1}{2}(n-3)}e^{-\frac{1}{2}ns^2}, \quad (2)$$

which should be compared with that of **70** (4).

This distribution is called Helmert's distribution, after the German astronomer and geodetist F. R. Helmert, who published it in 1876.

By expanding the m.g.f. and noting the coefficient of $\alpha^2/2!$ we find that the mean value of s^2 over all samples of n is $(n-1)\sigma^2/n$, where σ^2 is the variance of x. This

is really the theorem of mean square residual of **59** (5), and it is true not merely for normal but for general populations. Because of the factor $(n-1)/n$ the precept is often given to estimate variance by dividing the sum of squared deviations from m not by n but by $n-1$. On the other hand, the discrepancy in s^2 caused by not doing this is of order $1/n$, whereas the standard error of sampling of s^2 is of order $\sqrt{(2/n)}$. Thus the error of method is to the error of sampling in approximate ratio $1:\sqrt{(2n)}$, which even for n as small as 25 is less than 1/7. To insist on the divisor $n-1$ rather than n in large samples may therefore seem a little pedantic; but in small samples an appreciable difference is made. One advantage of the division by $n-1$ is this, that with the modified s^2 the probability function (2) assumes the form

$$2^{-\frac{1}{2}(n-1)}\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{-1}(n-1)^{\frac{1}{2}(n-1)}(s^2)^{\frac{1}{2}(n-3)}e^{-\frac{1}{2}(n-1)s^2}, \quad (3)$$

which is now of exactly the same form as in **70** (4), with $n-1$ for n throughout. Thus the loss of a degree of freedom is made apparent. In unstandardized units we must write s^2/σ^2 for s^2 and insert on the right of (3) the factor $1/\sigma^2$.

The m.g.f. of s^2 in (3) is $[1-2a/(n-1)]^{-\frac{1}{2}(n-1)}$, from which it follows, as in **70** (5), that the sampling variance of this modified s^2 is $2\sigma^4/(n-1)$, the standard error thus being $\sigma^2\sqrt{2}/\sqrt{(n-1)}$.

Example. By considering the coefficients of $a^3/3!$ and $a^4/4!$ in the s.g.f. investigate the skewness and excess of the distribution of s^2.

72. "Student's Ratio" t and its Distribution. We have seen in **69** that the mean m of a sample of n values x_j drawn from a normal population of mean μ and variance σ^2 is distributed normally with mean μ and variance σ^2/n. It follows that the standardized deviation $(m-\mu)\sqrt{n}/\sigma$ is distributed normally with mean zero and variance 1.

Now in practice we do not know σ^2 and so we cannot standardize the scale. All that we know is the estimate (taking $n-1$ for divisor) $s^2 = \Sigma(x_j - m)^2/(n-1)$. The deviation of the mean of sample from true mean, standardized by this estimate s^2, is thus $(m-\mu)\sqrt{n}/s = t$. This is " Student's Ratio," and it is not normally distributed.

" Student " was the pen-name under which W. S. Gosset (1876-1937) wrote his statistical papers. He discovered the distribution in 1908.

To simplify the distribution we may place the origin of x at $x = \mu$, thus putting $\mu = 0$. Then $m\sqrt{n}/s = t$. Since $m\sqrt{n}/s = (m\sqrt{n}/\sigma)/(s/\sigma)$, and since the distributions of $m\sqrt{n}/\sigma$ and s^2/σ^2 are independent of σ, we may use standard scale with $\sigma = 1$.

For *constant* s^2 we have $dm = sdt/\sqrt{n}$; also the probability of obtaining the value t is the probability that m takes the value $st/\sqrt{n}$, and the probability differential for this is

$$ce^{-\frac{1}{2}nm^2}dm = cn^{-\frac{1}{2}}se^{-\frac{1}{2}s^2t^2}dt. \quad . \quad . \quad (1)$$

This is for constant s^2; and so the probability differential of t is the integral of (1) over *all* values of s^2. Hence, multiplying (1) by the probability differential of s^2, which we already know from **71** (3), and integrating from 0 to ∞, we have

$$\begin{aligned} dp(t) &= c_1 dt \int_0^\infty se^{-\frac{1}{2}s^2t^2}s^{n-3}e^{-\frac{1}{2}(n-1)s^2}ds^2 \\ &= c_2[1+t^2/(n-1)]^{-\frac{1}{2}n}dt \quad . \quad . \quad (2) \end{aligned}$$

where

$$c_2 = \Gamma\left(\frac{n}{2}\right) \Big/ \left[(n-1)^{\frac{1}{2}}\Gamma(\tfrac{1}{2})\Gamma\left(\frac{n-1}{2}\right)\right] \quad . \quad (3)$$

the constant c_2 being fixed, as always, by the condition that the total probability is 1.

Note. The above derivation is the one usually given, but an important remark must be made. The essential step, the compounding of the probability differentials of m and s^2,

presumes the statistical independence of m and s^2. This independence (Appendix, **5**) is not evident, nor is it capable of quite elementary proof. The reader may assume, however, both here and in the case of the difference of means of two samples, that the numerator and denominator of t are independent.

The remarkable and important fact about the t-distribution is that it does not involve the unknown σ^2, a partial reason being that t is a *ratio*, of zero dimension in σ^2. The discovery of the distribution in 1908 had a profound influence on "small sample" theory; for whereas it had long been conventional to take s as the presumptive σ and to estimate the probable region of the unknown μ by regarding $(m-\mu)\sqrt{n}/s$ as a standardized *normal* variate, this was now seen to be an inexact procedure, and the t-distribution was used instead.

Since $[1+t^2/(n-1)]^{-\frac{1}{2}n}$ tends with increasing n to $\exp(-\frac{1}{2}t^2)$, it is apparent that for large samples the t-distribution tends to the standard normal one; but the tendency is not rapid, and for small values of n, as one might suspect from noting that $n=2$ gives the Cauchy distribution, the departure from normality is marked, the curves being platykurtic. For example, whereas in the normal curve 0·95 of the area is contained in the range $x=-1{\cdot}96$ to $x=1{\cdot}96$, in the t-curve for $n=10$ the same area lies between $t=-2{\cdot}26$ and $t=2{\cdot}26$; and for area 0·99 the ranges are given by $x=\pm 2{\cdot}58$ and $t=\pm 3{\cdot}25$.

A table of the probability integral of the t-distribution, in a form useful for practical application, is given in R. A. Fisher's *Statistical Methods for Research Workers*, 8th edition, p. 167. His n is our $n-1$, the number of degrees of freedom.

Example. A coin, thrown 20 times on each of 10 occasions, shows 7, 9, 6, 10, 13, 6, 9, 7, 10, 7 heads respectively. Assuming the binomial distribution of 20 throws to be approximately normal, consider whether the coin is biassed.

The mean of the heads thrown is $m=8{\cdot}4$ and $s^2=4{\cdot}93$, $s=2{\cdot}22$. Thus, presuming an unbiassed $\mu=10$, we have

$t = (10-8{\cdot}4)\sqrt{10}/2{\cdot}22 = 2{\cdot}28$. From Fisher's tables, in the row $n = 9$ (our $n-1$) we find $t = 2{\cdot}262$ at $P = 0{\cdot}05$. (P is the probability of a t *numerically* greater than 2·262.) Thus the coin-throws leave it rather doubtful whether the coin is biassed or not.

A reading of tables of the normal probability integral for $x = 2{\cdot}28$ would have given $P = 0{\cdot}023$, with an unjustifiably stronger suggestion of bias in the coin.

73. Difference of Means of Two Normal Samples. A valuable use of the t-distribution is in testing the hypothesis that two samples, with different numbers n and N in sample, are from the same normal population, of mean $\mu = 0$ and variance σ^2.

Let $x_1, x_2, \ldots, x_n$ be the first sample, $X_1, X_2, \ldots, X_N$ be the second, with respective means m and M, and estimates of variance

$$s^2 = \Sigma(x_j-m)^2/(n-1), \qquad S^2 = \Sigma(X_j-M)^2/(N-1). \quad (1)$$

The basis of the test is the difference $m-M$, the variance of which (**15**) is

$$\sigma^2/n+\sigma^2/N = (n+N)\sigma^2/nN. \quad . \quad . \quad (2)$$

The estimates s^2 and S^2 of σ^2 are (**71**) of weights $n-1$ and $N-1$, and so yield a combined estimate of σ^2, namely

$$\begin{aligned} \mathrm{s}^2 &= [(n-1)s^2+(N-1)S^2]/(n+N-2) \\ &= [\Sigma(x_j-m)^2+\Sigma(X_j-M)^2]/(n+N-2). \quad . \quad (3) \end{aligned}$$

It can be proved that $m-M$ and s^2 are statistically independent. We therefore define, from (2) and (3),

$$t = \frac{m-M}{\mathrm{s}}\sqrt{\left(\frac{nN}{n+N}\right)}, \quad . \quad . \quad . \quad (4)$$

and it now follows, by the argument of **72**, that this t has the t-distribution, but with $n+N-2$, the number of degrees of freedom used in estimating σ^2, in place of the former $n-1$. Thus the t-tables may be consulted for the probability $P(t)$ that t numerically exceeds any assigned value. (The examples of p. 109 are amenable to t-tests.)

The important point is the way in which σ^2 is estimated. One might have pooled both samples and estimated σ^2 from the squared $n+N$ deviations about the pooled mean $(nm+NM)/(n+N)$, summed and divided by $n+N-1$. This slightly more accurate estimate of σ^2 is, however, not independent of $m-M$.

74. The Ratio of Two Variates of the Same χ^2 Type. The two samples in **73** give in general different estimates s^2 and S^2 of the variance σ^2. If the question is whether both samples are from the same normal population, we shall wish to test this by means of s^2 and S^2, without reference to the unknown σ^2. The analogy of t suggests the ratio $u = s^2/S^2$. Since u is unaltered when we write s^2/σ^2 for s^2, and S^2/σ^2 for S^2, we may work in standard scale, using the s^2-distribution of **71**. Let us write $v = s^2/\sigma^2$, $V = S^2/\sigma^2$. Then $u = v/V$, or $v = uV$.

By **71** the probability differentials of v and V are

$$c_1 v^{\frac{1}{2}(n-3)} e^{-\frac{1}{2}(n-1)v} dv \text{ and } C_1 V^{\frac{1}{2}(N-3)} e^{-\frac{1}{2}(N-1)V} dV. \quad . \quad (1)$$

For *fixed* V we have $dv = V du$; so, integrating for *all* V, we have the probability differential of u,

$$\begin{aligned} & c_1 C_1 du \int_0^\infty V (uV)^{\frac{1}{2}(n-3)} e^{-\frac{1}{2}(n-1)uV} V^{\frac{1}{2}(N-3)} e^{-\frac{1}{2}(N-1)V} dV \\ &= c_1 C_1 u^{\frac{1}{2}(n-3)} du \int_0^\infty V^{\frac{1}{2}(n+N-4)} e^{-\frac{1}{2}(N-1+\overline{n-1}u)V} dV \\ &= c u^{\frac{1}{2}(n-3)} du / (N-1+\overline{n-1}\, u)^{\frac{1}{2}(n+N-2)}, \quad . \quad . \quad . \quad (2) \end{aligned}$$

where c is fixed by making the integral of u unity.

The distribution of u is thus given by (2). It is interesting to verify that as $N \to \infty$ the distribution tends to the χ^2 type, while if $n = 2$ we have a t^2 distribution.

The z-Distribution of Fisher. R. A. Fisher, in testing the difference of two estimates s^2 and S^2, uses not this ratio u but half its natural logarithm. If we put

$$z = \tfrac{1}{2} \log_e u, \quad u = e^{2z}, \quad du = 2e^{2z} dz,$$

the probability differential (2) becomes

$$dp = c_2 e^{(n-1)z} dz / [N-1+(n-1)e^{2z}]^{\frac{1}{2}(n+N-2)}, \quad . \quad (3)$$

where c_2 is such that the total integral of z is unity. The distribution thus obtained is Fisher's z-distribution.

Tables of $P(z)$, the probability of a z greater than an assigned value, are given in Fisher's *Statistical Methods for Research Workers*. In these tables the numerator of u is the greater of s^2 and S^2, so that z is positive; and the functions tabled are the values of z for assigned n and N, such that $P = 0{\cdot}05$, $0{\cdot}01$ and $0{\cdot}001$ respectively. The table for $P = 0{\cdot}001$ is due to C. G. Colcord and L. S. Deming.

75. Analysis of Variance and of Sum of Squares. The basic idea of the experimental designs introduced by R. A. Fisher, and of the accompanying technique called *analysis of variance*, is that of dividing up a total sum of squared deviations of a variate from its sample mean into several distinct sums of squares, each corresponding to a source, real or suspected, of variation. These partial sums yield estimates of the variance from each source, and the z-test is applied to ascertain whether these estimates are compatible with each other and with the estimate of residual variance. If they are not so compatible, it is presumed that the sources have distinct effects, which are further analysed, for example by difference (**73**) of means.

The resolution into sums of squares is founded on the Lemma, noted in **52** in connexion with the correlation ratio, that if k sets of $n_1, n_2, \ldots, n_k$ observations, with respective means M_j and mean square deviations S_j^2, are pooled in an aggregate of $n = n_1+n_2+\ldots+n_k$ observations, with mean M and mean square deviations S^2, then

$$nS^2 = \sum_j n_j(S_j^2+c_j^2), \quad . \quad . \quad . \quad (1)$$

where $c_j = M-M_j$.

For illustration we shall consider an experiment based on repeated trials and designed to ascertain (i) whether h

varieties of a cereal are different in crop yield, (ii) whether k kinds of fertilizing treatment are different in their effect on the crop yield of the h varieties.

Consider first the case (i), the experiment on varieties alone. Suppose each of them planted in k similar plots, assigned in random positions in a field, and subjected to uniform cultivation. The hk yields y_{ij}, where i refers to variety, j to plot-number, may be arranged for analysis in a rectangular scheme of h rows and k columns, a row to each variety. For convenience in the algebra let us choose the origin of y_{ij} so that the sum or mean of all y_{ij} is zero.

Now consider the sum $\Sigma\Sigma_{i\ j}\, y_{ij}^2$ over all hk deviations. Let the means of rows (varieties) be $y_{10}, y_{20}, \ldots, y_{h0}$. Then by (1), remembering that the general mean is zero, we have

$$\Sigma\Sigma_{i\ j}\, y_{ij}^2 = \Sigma\Sigma_{i\ j}\,(y_{ij}-y_{i0})^2+k\Sigma_{i}\, y_{i0}^2 \quad . \quad . \quad (2)$$

The sums here are sums of squared residuals, and under the assumption that all plot-yields have zero mean and variance σ^2, the mean values or expectations of the terms give, by **59** (5), the relation

$$(hk-1)\sigma^2 = h(k-1)\sigma^2+(h-1)\sigma^2, \quad . \quad (3)$$

where the terms correspond to those in (2). The first term on the right follows from the fact that the mean value or expectation of sum of the k squared deviations for any row is $(k-1)\sigma^2$; and the second term then follows by subtraction.

The coefficients in (3) are really *degrees of freedom*; and we thus distinguish $hk-1$ degrees of freedom for all hk plots, of which $h-1$ are for variation between means of rows, that is, *between* varieties, and $h(k-1)$ are for variation about the particular variety means y_{i0}, that is, *within* varieties.

If the hypothesis to be tested is that varieties are not

essentially different in yield, this is the same as to suppose that variation between varieties is subject to the same cause as variation within varieties, that is, to ordinary randomness arising from soil heterogeneity and other causes common to all plots. The test is therefore to compute an s^2 from the sum of squares between varieties and an S^2 from the sum of squares within varieties, these being independent estimates of σ^2, and to see from the z-table whether they are compatible. In the calculation of s^2 and S^2 the respective degrees of freedom should be used as divisors; and S^2 is most easily calculated by means of

$$h(k-1)S^2 = \Sigma\Sigma_{i\ j} y_{ij}^2 - k\Sigma_i y_{i0}^2. \quad . \quad . \quad . \quad (4)$$

76. Analysis into Two Sources of Variation and Residual. Next, still with the same h-by-k arrangement (which in the random placing of plots in the field is called the "randomized block" arrangement), let the rectangle of h rows and k columns of yields be set out for analysis in the case when there are not only h different varieties, but each is subjected to k different treatments, so that y_{ij} is the yield of the i^{th} variety under the j^{th} treatment. Let the means of columns (treatments) be $y_{01}, y_{02}, \ldots, y_{0k}$.

Consider the term $\Sigma\Sigma_{i\ j}(y_{ij}-y_{i0})^2$ in **75** (2), and imagine all the deviations from mean of variety, $y_{ij}-y_{i0}$, to be set out in a rectangle just as the y_{ij} were. Since $\Sigma_i y_{i0} = \Sigma\Sigma_{i\ j} y_{ij}/k = 0$, the means of the $y_{ij}-y_{i0}$ in *columns* are merely those of the y_{ij} themselves, namely $y_{01}, y_{02}, \ldots, y_{0k}$.

Hence, by analysing this term exactly as $\Sigma\Sigma_{i\ j} y_{ij}^2$ was, but with respect to *column* means instead of row means, we have

$$\Sigma\Sigma_{i\ j}(y_{ij}-y_{i0})^2 = \Sigma\Sigma_{i\ j}(y_{ij}-y_{i0}-y_{0j})^2 + h\Sigma_j y_{0j}^2. \quad . \quad (1)$$

Hence again, from 75 (2),

$$\Sigma\Sigma_{ij} y_{ij}^2 = \Sigma\Sigma_{ij}(y_{ij}-y_{i0}-y_{0j})^2+k\Sigma_i y_{i0}^2+h\Sigma_j y_{0j}^2, \quad . \quad (2)$$

which exhibits a threefold dissection, the last two terms on the right corresponding to variation between varieties and between treatments respectively, and the first term to residual variation. As for degrees of freedom, by taking expectations as before of these sums of squared residuals, we have

$$(hk-1)\sigma^2 = (h-1)(k-1)\sigma^2+(h-1)\sigma^2+(k-1)\sigma^2, \quad (3)$$

the coefficients giving the desired divisors of corresponding terms on the right of (2), for estimates of variance. The comparison of estimates by the z-table is then available.

77. The Latin Square. In the arrangement shown on the left, each of A, B, C, D, E appears exactly five times in rows and columns of a square, but no letter occurs twice in the same row or same column. Such an arrangement of h letters each repeated h times is called a *Latin square* of order h.

A	B	C	D	E
E	C	A	B	D
B	D	E	C	A
D	E	B	A	C
C	A	D	E	B

Imagine the Latin square to be a scheme of plot-yields set up for analysis, the letters representing yields of different varieties, the rows corresponding to varied treatments of one kind, the columns to varied treatments of another kind; for example, two kinds of fertilizer applied at once, at h different levels of strength in each. There are thus three dimensions of variation, two for treatments and one for variety; and so the yields may be written y_{ijl}, where i refers to row, j to column, l to variety. Let the respective means for rows, columns and varieties be y_{i00}, y_{0j0} and y_{00l}. Each suffix runs from 1 to h.

A first analysis as in **76** (2) gives us

$$\underset{ij}{\Sigma\Sigma} y_{ijl}^2 = \underset{ij}{\Sigma\Sigma}(y_{ijl}-y_{i00}-y_{0j0})^2+h\underset{i}{\Sigma} y_{i00}^2+h\underset{j}{\Sigma} y_{0j0}^2. \qquad (1)$$

But now arrange the $y_{ijl}-y_{i00}-y_{0j0}$ in rows, say, according to $l = 1, 2, \ldots, h$, and analyse once again. Since the means of y_{i00} and y_{0j0} are zero, the means of $y_{ijl}-y_{i00}-y_{0j0}$ are simply y_{00l}, where $l = 1, 2, \ldots, h$. We therefore have

$$\underset{ij}{\Sigma\Sigma} y_{ijl}^2 = \underset{ij}{\Sigma\Sigma}(y_{ijl}-y_{i00}-y_{0j0}-y_{00l})^2+h\underset{i}{\Sigma} y_{i00}^2+h\underset{j}{\Sigma} y_{0j0}^2+h\underset{i}{\Sigma} y_{00l}^2 \qquad (2)$$

where the three last terms on the right are sums of squares for variation between rows, columns and varieties respectively, and the first term is for residual variation. By taking the expectations of these sums of squared residuals we have

$$(h^2-1)\sigma^2 = (h-1)(h-2)\sigma^2+(h-1)\sigma^2+(h-1)\sigma^2+(h-1)\sigma^2, \qquad (3)$$

which shows the respective degrees of freedom to be used as divisors in the estimates of variance.

Example. The entries in the square below are the numbers of successes in 25 sets of 10 drawings with probability $p = 0{\cdot}52$, written down consecutively in 5 rows. The mean squares of the analyses may be compared with the theoretical σ^2, which is $10 \times 0{\cdot}52 \times 0{\cdot}48 = 2{\cdot}50$.

The working details, based on the formulæ of **75, 76, 77**, are shown (i) in ordinary row and column analysis, applicable equally to h rows and k columns, (ii) in Latin square analysis, using the particular Latin square (*q.v.*) given above.

(i)

						Sums.	Means.
	5	3	2	4	6	20	4·0
	6	7	5	4	5	27	5·4
	3	6	3	6	5	23	4·6
	8	3	7	6	4	28	5·6
	6	2	6	4	5	23	4·6
Sums	28	21	23	24	25	121	
Means	5·6	4·2	4·6	4·8	5·0		4·84

	Latin sq.	Sums.	Means.
(ii)	A	23	4·6
	B	22	4·4
	C	25	5·0
	D	29	5·8
	E	22	4·4
	Total .	121	4·84

The various sums of squares used are: (1) the sum of squares of all 25 entries, namely 647; (2) the sum of the five products, row-sum by row-mean, 594·2; (3) the same for columns, 591·0; (4) the same for letters in Latin square, 592·6. Each one of these must be corrected for transference to the general mean 4·84, and the correction in every case is to subtract the product of total sum by total mean, 121 by 4·84, or 585·64.

Thus the corrected sums of squares are (1) 61·36, (2) 8·56, (3) 5·36, (4) 6·96. The residual sum of squares is found by subtraction from the total sum, and the details of estimate of mean square are set out in tabular form thus:

(i) Row and column analysis.

	Sum sq.	Degr.	Mean sq.
Rows . .	8·56	4	2·14
Cols. . .	5·36	4	1·34
Res. . . .	47·44	16	2·97
Total . .	61·36	24	2·56

(ii) Latin square analysis.

	Sum sq.	Degr.	Mean sq.
Rows . .	8·56	4	2·14
Cols. . . .	5·36	4	1·34
Letters . .	6·96	4	1·74
Res. . . .	40·48	12	3·37
Total	61·36	24	2·56

We need not continue, but in practice the mean squares for rows, columns and letters would be compared with each other and with the residual mean square by taking half the

difference of logarithms and applying the z-test. (The logarithms are Napierian, and we may note the relation $\frac{1}{2}\log_e u = 1{\cdot}151\log_{10} u$.)

The principle of isolation, by appropriate experimental design, of the separate variations due to several simultaneous causes, has been developed and widely applied in recent years. Complex patterns, such as randomized blocks in which each element is itself a block, or Latin squares in which each "letter" is a Latin square, have been designed and used. The idea is to save time, space and expense by being able to conduct several kinds of experiment at the same time and within the one frame. For further details the reader may consult Fisher's *The Design of Experiments*, 2nd edition, or Yates's *The Design and Analysis of Factorial Experiments* (Harpenden, 1937).

78. Conclusion. The consideration of other sampling distributions would exceed our space and scope, but one of special interest may be noted. The distribution of r, the standardized product-moment estimate (without Sheppard's correction) of ρ in normal correlation, was found by R. A. Fisher in 1915. The probability function has the rather complicated form

$$\phi(r) = c(1-\rho^2)^{\frac{1}{4}(n-1)}(1-r^2)^{\frac{1}{4}(n-4)}\frac{d^{n-2}}{d(r\rho)^{n-2}}\left(\frac{\arccos(-r\rho)}{\sqrt{(1-r^2\rho^2)}}\right) \quad (1)$$

and the curve, if ρ is at all large and the sample small, is skew and in cases even U-shaped. (The function and its integral have been computed, for $n = 3, 4, \ldots, 25, 50, 100, 200, 400$ and $\rho = 0{\cdot}1, 0{\cdot}2, 0{\cdot}3, \ldots, 0{\cdot}9$, by F. N. David in *Tables of the Correlation Coefficient*, London, 1938.)

It was proved by Fisher (*Metron*, 1921) that the hyperbolic tangent transformation

$$z' = \tfrac{1}{2}\log\left(\frac{1+r}{1-r}\right), \quad \zeta = \tfrac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) \quad . \quad (2)$$

produces a distribution which even for n as small as 20 is nearly normal, with mean ζ and variance $1/(n-3)$.

A second transformation of r, namely

$$t = \frac{}{\sqrt{(1-r^2)}}\sqrt{n-2}, \qquad (3)$$

leads to a t-distribution with $n-2$ degrees of freedom. These transformations are necessary, because of the extreme non-normality of the sampling distribution of r, which makes the crude use of the standard error of r a fallacious procedure.

79. Estimation of Parameters from Sample. In **40** and **42** we have estimated the mean μ of a normal and a Poisson distribution by the mean m of the sample, in **43** we have pointed out the demerits of the mean of sample in estimating the true mean of a Cauchy distribution. The general problem of estimation is this: given n sample values $x_1, x_2, \ldots, x_n$ of a variate x with probability function $\phi(x\,;\theta)$ involving a parameter θ, what function $T(x_1, x_2, \ldots, x_n)$ of the sample values shall be used to estimate θ? The problem must be posed in mathematical terms, and must, in order to become intelligible, assume a certain degree of arbitrariness. One fruitful principle, well justified by its results, consists in choosing T by making the *compound probability density of* $x_1, x_2, \ldots, x_n$ a *maximum* with respect to θ. This is R. A. Fisher's *principle of maximum likelihood*. Another principle postulates (i) that T shall be *unbiassed*, in the sense that the mean value of T over all samples of n data shall be equal to θ, (ii) that of all such functions T shall be the one with *minimum sampling variance*. In many cases these two different approaches (the second of which has not yet been deeply explored) lead to the same function T of estimate. The situation is parallel to that which occurs in the theory of Least Squares, where, as mentioned at the end of **57**, different sets of postulates lead to the same normal equations.

80. Four-Place Table of $\Phi(x) = (2\pi)^{-\frac{1}{2}} \int_0^x e^{-\frac{1}{2}x^2} dx$.

x	Φ	x	Φ	x	Φ	x	Φ
0·00	0000	0·50	1915	1·00	3413	1·50	4332
0·02	0080	0·52	1985	1·02	3461	1·55	4394
0·04	0160	0·54	2054	1·04	3508	1·60	4452
0·06	0239	0·56	2123	1·06	3554	1·65	4505
0·08	0319	0·58	2190	1·08	3599	1·70	4554
0·10	0398	0·60	2257	1·10	3643	1·75	4599
0·12	0478	0·62	2324	1·12	3686	1·80	4641
0·14	0557	0·64	2389	1·14	3729	1·85	4678
0·16	0636	0·66	2454	1·16	3770	1·90	4713
0·18	0714	0·68	2517	1·18	3810	1·95	4744
0·20	0793	0·70	2580	1·20	3849	2·00	4772
0·22	0871	0·72	2642	1·22	3888	2·10	4821
0·24	0948	0·74	2703	1·24	3925	2·20	4861
0·26	1026	0·76	2764	1·26	3962	2·30	4893
0·28	1103	0·78	2823	1·28	3997	2·40	4918
0·30	1179	0·80	2881	1·30	4032	2·50	4938
0·32	1255	0·82	2939	1·32	4066	2·60	4953
0·34	1331	0·84	2995	1·34	4099	2·70	4965
0·36	1406	0·86	3051	1·36	4131	2·80	4974
0·38	1480	0·88	3106	1·38	4162	2·90	4981
0·40	1554	0·90	3159	1·40	4192	3·00	49865
0·42	1628	0·92	3212	1·42	4222	3·20	49931
0·44	1700	0·94	3264	1·44	4251	3·40	49966
0·46	1772	0·96	3315	1·46	4279	3·60	49984
0·48	1844	0·98	3365	1·48	4306	3·80	49993
0·50	1915	1·00	3413	1·50	4332	4·00	49997

A decimal point is understood before each entry $\Phi(x)$; and second difference interpolation is advisable in the last column.

A useful inverse table of the normal probability integral is the table of Probits in Fisher and Yates's *Statistical Tables for Biological, Agricultural and Medical Research* (Oliver and Boyd, 1938), pp. 38-40. The " probit " is the value of x which cuts off at its ordinate a given percentage of area measured from the left of the normal curve.

APPENDIX

1. Finite Differences and Factorial Polynomials. Most tables of functions provide us with sequences of values which by a suitable choice of origin and scale may be denoted by $u_0, u_1, u_2, \ldots$. To these we may apply differencing and repeated differencing, analogous in the Calculus of Finite Differences to differentiation in the Infinitesimal Calculus (see Whittaker and Robinson, *Calculus of Observations* (Blackie), Chapters I to IV). The operations most commonly used are :

$$\begin{aligned}
&\text{the advancing difference,} && \Delta u_x = u_{x+1} - u_x, \\
&\text{the receding difference,} && \nabla u_x = u_x - u_{x-1}, \\
&\text{the central difference,} && \delta u_x = u_{x+\frac{1}{2}} - u_{x-\frac{1}{2}}, \\
&\text{the averaging operation,} && \mu u_x = \tfrac{1}{2}(u_{x+\frac{1}{2}} + u_{x-\frac{1}{2}}), \\
&\text{the mean central difference} && \mu\delta u_x = \tfrac{1}{2}(u_{x+1} - u_{x-1}),
\end{aligned} \qquad (1)$$

operations which may all be repeated. The classical formula of interpolation, which uses advancing differences derived from $u_0, u_1, u_2 \ldots$, is the Gregory-Newton formula

$$u_x = u_0 + x\Delta u_0 + x^{(2)}\Delta^2 u_0/2! + x^{(3)}\Delta^3 u_0/3! + \ldots, \qquad (2)$$

a formula which terminates at $n+1$ terms if u_x is a polynomial of degree n, and which in practical cases converges well, with negligible remainder, after a few terms. The formula is the analogue of the Taylor series in the Infinitesimal Calculus.

The polynomials $1, x, x^{(2)}, x^{(3)}, \ldots$ which appear in (2) are (**29**) ordinary *factorial* polynomials $1, x, x(x-1), x(x-1)(x-2), \ldots$. If they are divided respectively by 0!, 1!, 2!, 3!, .. we obtain the *reduced factorials* or binomial coefficients $1, x, x_{(2)}, x_{(3)}, \ldots$.

Central factorials may be defined by

$$1, x^{\{1\}} = x, x^{\{2\}} = (x+\tfrac{1}{2})(x-\tfrac{1}{2}), x^{\{3\}} = (x+1)(x(x-1), \ldots, \qquad (3)$$

the factors being in arithmetical progression of common difference unity and centred at x. The reader may verify that

$$\mu x^{\{1\}} = x, \ \mu x^{\{2\}} = x^2, \ \mu x^{\{3\}} = x(x^2 - \tfrac{1}{4}), \ \mu x^{\{4\}} = x^2(x^2 - 1), \qquad (4)$$

Given an odd number of values of u_x with central value u_0, the Newton-Stirling formula of interpolation is useful,

$$u_x = u_0 + x \cdot \mu\delta u_0 + \mu x^{\{2\}} \cdot \delta^2 u_0/2! + x^{\{3\}} \cdot \mu\delta^3 u_0/3! + \dots . \quad (5)$$

Given an even number of values with two central values $u_{-\frac{1}{2}}$ and $u_{\frac{1}{2}}$, the Newton-Bessel formula is the appropriate one,

$$u_x = \mu u_0 + \mu x \cdot \delta u_0 + x^{\{2\}} \mu\delta^2 u_0/2! + \mu x^{\{3\}} \cdot \delta^2 u_0/3! + \dots . \quad (6)$$

These formulæ use central and mean central factorials, and mean central and central differences, alternately.

The origin of interpolation $x = 0$ can almost always be chosen so that x, in the interpoland u_x, need not exceed $\frac{1}{4}$.

The following relations are fundamental:

$$\Delta x^{(r)} = r x^{(r-1)}, \text{ equivalent to } \Delta x_{(r)} = x_{(r-1)}, \ \delta x^{\{r\}} = r x^{\{r-1\}}.$$

(Cf. $Dx^r = rx^{r-1}$ in the Differential Calculus.)

2. Finite Sums. The following table of repeated summation upon $u_0, u_1, \dots, u_{n-1}$, exemplified for $n = 5$, follows the scheme proposed in **19** for computing factorial moments.

u	Σ	Σ^2	Σ^3	Σ^4	Σ^5
u_0	$u_0+u_1+u_2+u_3+u_4$				
u_1	$u_1+u_2+u_3+u_4$	$u_1+2u_2+3u_3+4u_4$			
u_2	$u_2+u_3+u_4$	$u_2+2u_3+3u_4$	$u_2+3u_3+6u_4$		
u_3	u_3+u_4	u_3+2u_4	u_3+3u_4	u_3+4u_4	
u_4	u_4	u_4	u_4	u_4	u_4

Scrutiny will show that the entries at the tops of the successive columns of summation are the reduced factorial moments:

$$\Sigma u_x, \ \Sigma x u_x, \ \Sigma x_{(2)} u_x, \ \Sigma x_{(3)} u_x, \dots .$$

This may be proved by an induction based on $\Delta x_{(r)} = x_{(r-1)}$.

With a little more difficulty, using central factorials, it may be proved that the scheme of repeated summation toward the centre with alternate averaging, used in Ex. 3 of **19**, produces *reduced* central and mean central factorial moments $\Sigma x^{(r)} u_x/r!$ and $\Sigma \mu x^{(r)} u_x/r!$.

3. Relations between Powers and Factorials. We have

(i) $x = x,$
$x^2 = x^{(2)} + x,$
$x^3 = x^{(3)} + 3x^{(2)} + x,$
$x^4 = x^{(4)} + 6x^{(3)} + 7x^{(2)} + x,$

(ii) $x = x$
$x^2 = \mu x^{\{2\}},$
$x^3 = x^{\{3\}} + x,$
$x^4 = \mu x^{\{4\}} + \mu x^{\{2\}}.$

as may be verified by actual expansion. Multiplying any of these relations by u_x and summing over equally spaced values of x, (i) with $x = 0$ as least value, (ii) with $x = 0$ as middle value, we derive the relations quoted and used in **19**, Exs. 1 and 3, for converting factorial moments, or central and mean central factorial moments, into ordinary moments.

4. Tables of Normal Probability Integral and Poisson Function. A very convenient table of the normal probability integral in standard scale, to four places of decimals, is given in Bowley's *Elements of Statistics*, p. 271. The table is accurate enough for most practical purposes, and may be interpolated by proportional parts, that is, only using first differences. We give a compact table in **80**, p. 144.

In the Poisson function the chief requirement is the value of e^{-m}. If a machine is available, the following short table enables e^{-m} to be computed with sufficient accuracy for $m = 0$ to 10.

m	e^{-m}	m	e^{-m}	m	e^{-m}	m	e^{-m}
1	0·36788	0·1	0·90484	0·01	0·99005	0·001	0·99900
2	0·13534	0·2	0·81873	0·02	0·98020	0·002	0·99800
3	0·049787	0·3	0·74082	0·03	0·97045	0·003	0·99700
4	0·018316	0·4	0·67032	0·04	0·96079	0·004	0·99601
5	0·0067379	0·5	0·60653	0·05	0·95123	0·005	0·99501
6	0·0024788	0·6	0·54881	0·06	0·94176	0·006	0·99402
7	0·0009119	0·7	0·49659	0·07	0·93239	0·007	0·99302
8	0·0003355	0·8	0·44933	0·08	0·92312	0·008	0·99203
9	0·0001234	0·9	0·40657	0·09	0·91393	0·009	0·99104
10	0·0000454	1·0	0·36788	0·10	0·90484	0·010	0·99005

For smaller values of m than those given above the approximation $1-m$ for e^{-m} is correct to at least five decimals.

Ex. In the example of **42** we have $m = 3{\cdot}870$. Entering the above table at $m = 3$, 0·8, 0·07, we form the product, thus: $0{\cdot}049787 \times 0{\cdot}44933 \times 0{\cdot}93239 = 0{\cdot}20838.$

5. Linear Dependence, Functional Dependence, Correlation, Statistical Dependence. These are concepts which need careful discrimination. The functions $u_j(x)$, where $j = 1, 2, \ldots, n$, are *linearly dependent* if a relation

$$c_1u_1+c_2u_2+\ldots+c_nu_n = 0, \quad . \quad . \quad . \quad (1)$$

exists identically in x, where one at least of the c_j is not zero. They are *functionally dependent* if a functional relation

$$F(u_1, u_2, \ldots, u_n) = 0 \quad . \quad . \quad . \quad (2)$$

exists identically in x. Linear dependence, for example, is the case where F is a non-zero linear function. They are *uncorrelated* if the product moment μ_{11} vanishes for each pair u_i and u_j of the set.

Correlation and functional dependence are (48) not necessarily the same. The simplest example is perhaps $u = a \cos x+b \sin x$, $v = a \sin x-b \cos x$. Here u and v are uncorrelated, yet are dependent in view of the quadratic relation $u^2+v^2-a^2-b^2 = 0$.

To describe *statistical dependence*, we may say that statistical independence is really obedience to the multiplication theorem of probability. Suppose we have two functions of n variates, $u(x, y, z)$ and $v(x, y, z)$, where we illustrate by $n = 3$. They have each a probability, or probability density, let us say $\psi_1(u)$ and $\psi_2(v)$, depending on the distribution of x, y and z. They have also a compound probability, or probability density, let us say $\psi(u, v)$. If for all the possible values of x, y, z we have $\psi(u, v) = \psi_1(u)\psi_2(v)$, then we say that u and v are statistically independent.

An equivalent formulation is by generating functions. If

$$G(\alpha, \beta) = \iiint e^{\alpha u+\beta v}\phi(x, y, z)dxdydz \quad (3)$$

where $\phi(x, y, z)$ is the compound probability density of x, y, z, and if

$$G(\alpha, \beta) = G(\alpha, 0)G(0, \beta), \quad . \quad . \quad . \quad (4)$$

all integrals existing in some common domain of α, β, then u and v are statistically independent. By this criterion it may be proved that the estimates m of μ and s^2 of σ^2 in a normal sample (**72**) of n values x_j are statistically independent, so that the derivation of the t-distribution (*loc. cit.*) is valid.

INDEX

PRINTED IN GREAT BRITAIN BY
OLIVER AND BOYD LTD.
EDINBURGH